Universal Source Coding for Monotonic and Fast Decaying Monotonic Distributions[*]

Gil I. Shamir
Department of Electrical and Computer Engineering
University of Utah
Salt Lake City, UT 84112, U.S.A
e-mail: gshamir@ece.utah.edu.

Abstract

We study universal compression of sequences generated by monotonic distributions. We show that for a monotonic distribution over an alphabet of size k, each probability parameter costs essentially $0.5 \log(n/k^3)$ bits, where n is the coded sequence length, as long as $k = o(n^{1/3})$. Otherwise, for $k = O(n)$, the total average sequence redundancy is $O(n^{1/3+\varepsilon})$ bits overall. We then show that there exists a sub-class of monotonic distributions over infinite alphabets for which redundancy of $O(n^{1/3+\varepsilon})$ bits overall is still achievable. This class contains fast decaying distributions, including many distributions over the integers and geometric distributions. For some slower decays, including other distributions over the integers, redundancy of $o(n)$ bits overall is achievable, where a method to compute specific redundancy rates for such distributions is derived. The results are specifically true for finite entropy monotonic distributions. Finally, we study individual sequence redundancy behavior assuming a sequence is governed by a monotonic distribution. We show that for sequences whose empirical distributions are monotonic, individual redundancy bounds similar to those in the average case can be obtained. However, even if the monotonicity in the empirical distribution is violated, diminishing per symbol individual sequence redundancies with respect to the monotonic maximum likelihood description length may still be achievable.

Index Terms: monotonic distributions, universal compression, average redundancy, individual redundancy, large alphabets, patterns.

[*]Supported by NSF Grant CCF-0347969. Part of the material in this manuscript was accepted for presentation in the IEEE International Symposium on Information Theory, Nice, France, June, 2007.

1 Introduction

The classical setting of the universal lossless compression problem [5], [8], [9] assumes that a sequence x^n of length n that was generated by a source $\boldsymbol{\theta}$ is to be compressed without knowledge of the particular $\boldsymbol{\theta}$ that generated x^n but with knowledge of the class Λ of all possible sources $\boldsymbol{\theta}$. The average performance of any given code, that assigns a length function $L(\cdot)$, is judged on the basis of the *redundancy* function $R_n\left(L, \boldsymbol{\theta}\right)$, which is defined as the difference between the expected code length of $L\left(\cdot\right)$ with respect to (w.r.t.) the given source probability mass function P_θ and the nth-order entropy of P_θ normalized by the length n of the uncoded sequence. A class of sources is said to be universally compressible in some worst sense if the redundancy function diminishes for this worst setting. Another approach to universal coding [29] considers the *individual sequence* redundancy $\hat{R}_n\left(L, x^n\right)$, defined as the normalized difference between the code length obtained by $L(\cdot)$ for x^n and the negative logarithm of the *maximum likelihood* (ML) probability of the sequence x^n, where the ML probability is within the class Λ. We thereafter refer to this negative logarithm as the *ML description length* of x^n. The individual sequence redundancy is defined for each sequence that can be generated by a source $\boldsymbol{\theta}$ in the given class Λ.

Classical literature on universal compression [5], [8], [9], [23], [29] considered compression of sequences generated by sources over finite alphabets. In fact, it was shown by Kieffer [15] (see also [13]) that there are no universal codes (in the sense of diminishing redundancy) for sources over infinite alphabets. Later work (see, e.g., [21], [25]), however, bounded the achievable redundancies for *identically and independently distributed* (i.i.d.) sequences generated by sources over large and infinite alphabets. Specifically, while it was shown that the redundancy does not decay if the alphabet size is of the same order of magnitude as the sequence length n or greater, it was also shown that the redundancy does decay for alphabets of size $o(n)$. [1]

While there is no universal code for infinite alphabets, recent work [20] demonstrated that if one considers the *pattern* of a sequence instead of the sequence itself, universal codes do exist in the sense of diminishing redundancy. A pattern of a sequence, first considered, to the best of our knowledge, in [1], is a sequence of indices, where the index ψ_i at time i represents the order of first occurrence of letter x_i in the sequence x^n. Further study of universal compression of patterns [20], [21], [26], [28] provided various lower and upper bounds to various forms of redundancy in universal

[1] For two functions $f(n)$ and $g(n)$, $f(n) = o(g(n))$ if $\forall c, \exists n_0$, such that, $\forall n > n_0$, $f(n) < cg(n)$; $f(n) = O(g(n))$ if $\exists c, n_0$, such that, $\forall n > n_0$, $0 \leq f(n) \leq cg(n)$; $f(n) = \Theta(g(n))$ if $\exists c_1, c_2, n_0$, such that, $\forall n > n_0$, $c_1 g(n) \leq f(n) \leq c_2 g(n)$.

compression of patterns. Another related study is that of compression of data, where the order of the occurring data symbols is not important, but their types and empirical counts are [30]-[31].

This paper considers universal compression of data sequences generated by distributions that are known *a-priori* to be monotonic. Hence, the order of probabilities of the source symbols is known in advance to both encoder and decoder and can be utilized as side information to improve universal compression performance. Monotonic distributions are common for distributions over the integers, including the geometric distribution and others. Such distributions do occur in image compression problems (see, e.g., [18], [19]), and in other applications that compress residual signals. A specific application one can consider for the results in this paper is compression of the list of last or first names in a given city of a given population. One can usually find some monotonicity for such a distribution in the given population, which both encoder and decoder may be aware of *a-priori*. For example, the last name "Smith" can be expected to be much more common than the last name "Shannon". Another example is the compression of a sequence of observations of different species, where one has prior knowledge which species are more common, and which are rare. Finally, one can consider compressing data for which side information given to the decoder through a different channel gives the monotonicity order.

Unlike compression of patterns, Foster, Stine, and Wyner, showed in [10] that there are no universal block codes in the standard sense for the complete class of monotonic distributions. The main reason is that there exist such distributions, for which much of the statistical weight lies in symbols that have very low probability, and most of which will not occur in a given sequence. Thus, in practice, even though one has the prior knowledge of the monotonicity of the distribution, this monotonicity is not necessarily retained in an observed sequence. Therefore, actual coding can be very similar to compressing with infinite alphabets, and the additional prior knowledge of the monotonicity is not very helpful in reducing redundancy. Despite that, Foster, Stine, and Wyner demonstrated codes that obtained universal per-symbol redundancy of $o(1)$ as long as the source entropy is fixed (i.e., neither increasing with n nor infinite). However, instead of considering redundancy in the standard sense, the study of monotonic distributions resorted to studying *relative redundancy*, which bounds the ratio between average assigned code length and the source entropy. This approach dates back to work by Elias [7], Rissanen [22], and Ryabko [24].

The work in [10] studied coding sequences (or blocks) generated by i.i.d. monotonic distributions, and designed codes for which the relative block redundancy could be (upper) bounded. Unlike that work, the focus in [7], [22], and [24] was on designing codes that minimize the redundancy or relative redundancy for a single symbol generated by a monotonic distribution. Specifically, in [22], *minimax* codes, which minimize the relative redundancy for the worst possible monotonic distribution over a given alphabet size, were derived. In [24], it was shown that redundancy of $O(\log \log k)$, where k is the alphabet size, can be obtained with minimax per-symbol codes. Very recent work [16] considered per-symbol codes that minimize an average redundancy over the class of monotonic distributions for a given alphabet size. Unlike [10], all these papers study per-symbol codes. Therefore, the codes designed always pay non-diminishing per-symbol redundancy.

A different line of work on monotonic distributions considered optimizing codes for a known monotonic distribution but with unknown parameters (see [18], [19] for design of codes for two-sided geometric distributions). In this line of work, the class of sources is very limited and consists of only the unknown parameters of a known distribution.

In this paper, we consider a general class of monotonic distributions that is not restricted to a specific type. We study standard block redundancy for coding sequences generated by i.i.d. monotonic distributions, i.e., a setting similar to the work in [10]. We do, however, restrict ourselves to smaller subsets of the complete class of monotonic distributions. First, we consider monotonic distributions over alphabets of size k, where k is either small w.r.t. n, or of $O(n)$. Then, we extend the analysis to show that under minimal restrictions of the monotonic distribution class, there exist universal codes in the standard sense, i.e., with diminishing per-symbol redundancy. In fact, not only do universal codes exist, but under mild restrictions, they achieve the same redundancy as obtained for alphabets of size $O(n)$. The restrictions on this subclass imply that some types of fast decaying monotonic distributions are included in it, and therefore, sequences generated by these distributions (without prior knowledge of either the distribution or of its parameters) can still be compressed universally in the class of monotonic distributions.

The main contributions of this paper are the development of codes and derivation of their upper bounds on the redundancies for coding i.i.d. sequences generated by monotonic distributions. Specifically, the paper gives complete characterization of the redundancy in coding with monotonic distributions over "small" alphabets ($k = o(n^{1/3})$) and "large" alphabets ($k = O(n)$). Then, it shows that these redundancy bounds carry over (in first order) to fast decaying distributions. Next, a code that achieves good redundancy rates for even slower decaying monotonic distributions is derived, and is used to study achievable redundancy rates for such distributions. Lower bounds are also presented to complete the characterization, and are shown to meet the upper bounds in the first three cases (small alphabets, large alphabets, and fast decaying distributions). The lower bounds turn out to result from lower bounds obtained for coding patterns. The relationship to patterns is demonstrated in the proofs of those lower bounds. Finally, individual sequences are considered. It is shown that under mild conditions, there exist universal codes w.r.t. the monotonic ML description length for sequences that contain the $O(n)$ more likely symbols, even if their empirical distributions are not monotonic.

The outline of this paper is as follows. Section 2 describes the notation and basic definitions. Then, in section 3, lower bounds on the redundancy for monotonic distributions are derived. Next, in Section 4, we propose codes and upper bound their redundancy for coding monotonic distributions over small and large alphabets. These bounds are then extended to fast decaying monotonic distributions in Section 5. Finally, in Section 6, we consider individual sequence redundancy.

2 Notation and Definitions

Let $x^n \stackrel{\triangle}{=} (x_1, x_2, \ldots, x_n)$ denote a sequence of n symbols over the alphabet Σ of size k, where k can go to infinity. Without loss of generality, we assume that $\Sigma = \{1, 2, \ldots, k\}$, i.e., it is the set of positive integers from 1 to k. The sequence x^n is generated by an i.i.d. distribution of some source, determined by the parameter vector $\boldsymbol{\theta} \stackrel{\triangle}{=} (\theta_1, \theta_2, \ldots, \theta_k)$, where θ_i is the probability of X taking value i. The components of $\boldsymbol{\theta}$ are non-negative and sum to 1. The distributions we consider in this paper are monotonic. Therefore, $\theta_1 \geq \theta_2 \geq \ldots \geq \theta_k$. The class of all monotonic distributions will be denoted by $\mathcal{M}$. The class of monotonic distributions over an alphabet of size k is denoted by $\mathcal{M}_k$. It is assumed that prior to coding x^n both encoder and decoder know that $\boldsymbol{\theta} \in \mathcal{M}$ or $\boldsymbol{\theta} \in \mathcal{M}_k$, and also know the order of the probabilities in $\boldsymbol{\theta}$. In the more restrictive setting, k is known in advance and it is known that $\boldsymbol{\theta} \in \mathcal{M}_k$. We do not restrict ourselves to this setting. In general, boldface letters will denote vectors, whose components will be denoted by their indices in the vector. Capital letters will denote random variables. We will denote an estimator by the *hat* sign. In particular, $\hat{\boldsymbol{\theta}}$ will denote the ML estimator of $\boldsymbol{\theta}$ which is obtained from x^n.

The probability of x^n generated by $\boldsymbol{\theta}$ is given by $P_\theta(x^n) \stackrel{\triangle}{=} \Pr(x^n \mid \boldsymbol{\Theta} = \boldsymbol{\theta})$. The average per-symbol[2] nth-order redundancy obtained by a code that assigns length function $L(\cdot)$ for $\boldsymbol{\theta}$ is

$$R_n(L, \boldsymbol{\theta}) \stackrel{\triangle}{=} \frac{1}{n} E_\theta L[X^n] - H_\theta[X], \tag{1}$$

where E_θ denotes expectation w.r.t. $\boldsymbol{\theta}$, and $H_\theta[X]$ is the (per-symbol) entropy (rate) of the source

[2] In this paper, redundancy is defined per-symbol (normalized by the sequence length n). However, when we refer to redundancy in overall bits, we address the block redundancy cost for a sequence.

($H_\theta\left[X^n\right]$ is the nth-order sequence entropy of $\boldsymbol{\theta}$, and for i.i.d. sources, $H_\theta\left[X^n\right] = nH_\theta\left[X\right]$). With entropy coding techniques, assigning a universal probability $Q\left(x^n\right)$ is identical to designing a universal code for coding x^n where, up to negligible integer length constraints that will be ignored, the negative logarithm to the base of 2 of the assigned probability is considered as the code length.

The *individual* sequence redundancy (see, e.g., [29]) of a code with length function $L\left(\cdot\right)$ per sequence x^n is

$$\hat{R}_n\left(L, x^n\right) \stackrel{\triangle}{=} \frac{1}{n}\left\{L\left(x^n\right) + \log P_{ML}\left(x^n\right)\right\}, \tag{2}$$

where the logarithm function is taken to the base of 2, here and elsewhere, and $P_{ML}\left(x^n\right)$ is the probability of x^n given by the ML estimator $\hat{\boldsymbol{\theta}}_\Lambda \in \Lambda$ of the governing parameter vector $\boldsymbol{\Theta}$. The negative logarithm of this probability is, up to integer length constraints, the shortest possible code length assigned to x^n in Λ. It will be referred to as the *ML description length* of x^n in Λ. In the general case, one considers the i.i.d. ML. However, since we only consider $\boldsymbol{\theta} \in \mathcal{M}$, i.e., restrict the sequence to one governed by a monotonic distribution, we define $\hat{\boldsymbol{\theta}}_{\mathcal{M}} \in \mathcal{M}$ as the monotonic ML estimator. Its associated shortest code length will be referred to as the *monotonic ML description length*. The estimator $\hat{\boldsymbol{\theta}}_{\mathcal{M}}$ may differ from the i.i.d. ML $\hat{\boldsymbol{\theta}}$, in particular, if the empirical distribution of x^n is not monotonic. The individual sequence redundancy in $\mathcal{M}$ is thus defined w.r.t. the monotonic ML description length, which is the negative logarithm of $P_{ML}\left(x^n\right) \stackrel{\triangle}{=} P_{\hat{\theta}_{\mathcal{M}}}\left(x^n\right) \stackrel{\triangle}{=} \Pr\left(x^n \mid \boldsymbol{\Theta} = \hat{\boldsymbol{\theta}}_{\mathcal{M}} \in \mathcal{M}\right)$.

The average *minimax* redundancy of some class Λ is defined as

$$R_n^+\left(\Lambda\right) \stackrel{\triangle}{=} \min_L \sup_{\boldsymbol{\theta}\in\Lambda} R_n\left(L, \boldsymbol{\theta}\right). \tag{3}$$

Similarly, the *individual minimax* redundancy is that of the best code $L\left(\cdot\right)$ for the worst sequence x^n,

$$\hat{R}_n^+\left(\Lambda\right) \stackrel{\triangle}{=} \min_L \sup_{\boldsymbol{\theta}\in\Lambda} \max_{x^n} \frac{1}{n}\left\{L\left(x^n\right) + \log P_\theta\left(x^n\right)\right\}. \tag{4}$$

The *maximin* redundancy of Λ is

$$R_n^-\left(\Lambda\right) \stackrel{\triangle}{=} \sup_w \min_L \int_\Lambda w\left(d\boldsymbol{\theta}\right) R_n\left(L, \boldsymbol{\theta}\right), \tag{5}$$

where $w(\cdot)$ is a prior on Λ. In [5], it was shown that $R_n^+\left(\Lambda\right) \geq R_n^-\left(\Lambda\right)$. Later, however, [6], [11], [24] the two were shown to be essentially equal.

3 Lower Bounds

Lower bounds on various forms of the redundancy for the class of monotonic distributions can be obtained with slight modifications of the proofs for the lower bounds on the redundancy of coding patterns in [14], [20], [21], and [26]. The bounds are presented in the following three theorems. For the sake of completeness, the main steps of the proofs of the first two theorems are presented in appendices, and the proof of the third theorem is presented below. The reader is referred to [14], [20], [21], [25] and [26] for more details.

Theorem 1 *Fix an arbitrarily small $\varepsilon > 0$, and let $n \to \infty$. Then, the nth-order average maximin and minimax universal coding redundancies for i.i.d. sequences generated by a monotonic distribution with alphabet size k are lower bounded by*

$$R_n^-\left(\mathcal{M}_k\right) \geq \begin{cases} \frac{k-1}{2n}\log\frac{n^{1-\varepsilon}}{k^3} + \frac{k-1}{2n}\log\frac{\pi e^3}{2} - O\left(\frac{\log k}{n}\right), & \textit{for } k \leq \left(\frac{\pi n^{1-\varepsilon}}{2}\right)^{1/3} \\ \left(\frac{\pi}{2}\right)^{1/3}\cdot\left(1.5\log e\right)\cdot\frac{n^{(1-\varepsilon)/3}}{n} - O\left(\frac{\log n}{n}\right), & \textit{for } k > \left(\frac{\pi n^{1-\varepsilon}}{2}\right)^{1/3} \end{cases}. \tag{6}$$

Theorem 2 *Fix an arbitrarily small $\varepsilon > 0$, and let $n \to \infty$. Then, the nth-order average universal coding redundancy for coding i.i.d. sequences generated by monotonic distributions with alphabet size k is lower bounded by*

$$R_n\left(L, \boldsymbol{\theta}\right) \geq \begin{cases} \frac{k-1}{2n}\log\frac{n^{1-\varepsilon}}{k^3} - \frac{k-1}{2n}\log\frac{8\pi}{e^3} - O\left(\frac{\log k}{n}\right), & \textit{for } k \leq \frac{1}{2}\cdot\left(\frac{n^{1-\varepsilon}}{\pi}\right)^{1/3} \\ \frac{1.5\log e}{2\pi^{1/3}}\cdot\frac{n^{(1-\varepsilon)/3}}{n} - O\left(\frac{\log n}{n}\right), & \textit{for } k > \frac{1}{2}\cdot\left(\frac{n^{1-\varepsilon}}{\pi}\right)^{1/3} \end{cases} \tag{7}$$

for every code $L(\cdot)$ and almost every i.i.d. source $\boldsymbol{\theta} \in \mathcal{M}_k$, except for a set of sources $A_\varepsilon(n)$ whose relative volume in $\mathcal{M}_k$ goes to 0 as $n \to \infty$.

Theorems 1 and 2 give lower bounds on redundancies of coding over monotonic distributions for the class $\mathcal{M}_k$. However, the bounds are more general, and the second region applies to the whole class of monotonic distributions $\mathcal{M}$. As in the case of patterns [20], [26], the bounds in (6)-(7) show that each parameter costs at least $0.5\log(n/k^3)$ bits for small alphabets, and the total universality cost is at least $\Theta(n^{1/3-\varepsilon})$ bits overall for larger alphabets. Unlike the currently known results on patterns, however, we show in Section 4 that for $k = O(n)$ these bounds are achievable for monotonic distributions. The proofs of Theorems 1 and 2 are presented in Appendix A and in Appendix B, respectively.

Theorem 3 *Let $n \to \infty$. Then, the nth-order individual minimax redundancy for i.i.d. sequences with maximal letter k w.r.t. the monotonic ML description length with alphabet size k is lower bounded by*

$$\hat{R}_n^+ \left(\mathcal{M}_k\right) \geq \begin{cases} \frac{k-1}{2n} \log \frac{n}{k^3} + \frac{k}{n} \log \frac{e^{23/12}}{\sqrt{2\pi}} - O\left(\frac{\log k}{n}\right), & \textit{for } k \leq \frac{e^{5/18}}{(2\pi)^{1/3}} \cdot n^{1/3} \\ \frac{e^{5/18}}{(2\pi)^{1/3}} \cdot \frac{3}{2}(\log e) \cdot \frac{n^{1/3}}{n} - O\left(\frac{\log n}{n}\right), & \textit{for } n > k > \frac{e^{5/18}}{(2\pi)^{1/3}} \cdot n^{1/3} \\ \frac{3}{2}(\log e) \cdot \frac{n^{1/3}}{n} - O\left(\frac{\log n}{n}\right), & \textit{for } k \geq n. \end{cases} \tag{8}$$

Theorem 3 lower bounds the individual minimax redundancy for coding a sequence believed to have an empirical monotonic distribution. The alphabet size is determined by the maximal letter that occurs in the sequence, i.e., $k = \max\{x_1, x_2, \ldots, x_n\}$. (If k is unknown, one can use Elias' code for the integers [7] using $O(\log k)$ bits to describe k. However this is not reflected in the lower bound.) The ML probability estimate is taken over the class of monotonic distributions, i.e., the empirical probability (standard ML) estimate $\hat{\boldsymbol{\theta}}$ is not $\hat{\boldsymbol{\theta}}_{\mathcal{M}}$ in case $\hat{\boldsymbol{\theta}}$ does not satisfy the monotonicity that defines the class $\mathcal{M}$. While the average case maximin and minimax bounds of Theorem 1 also apply to $\hat{R}_n^+ \left(\mathcal{M}_k\right)$, the bounds of Theorem 3 are tighter for the individual redundancy and are obtained using individual sequence redundancy techniques.

Proof of Theorem 3: Using Shtarkov's *normalized maximum likelihood* (NML) approach [29], one can assign probability

$$Q\left(x^n\right) \stackrel{\triangle}{=} \frac{P_{\hat{\theta}_{\mathcal{M}}}\left(x^n\right)}{\sum_{y^n} P_{\hat{\theta}_{\mathcal{M}}}\left(y^n\right)} \stackrel{\triangle}{=} \frac{\max_{\theta' \in \mathcal{M}} P_{\theta'}\left(x^n\right)}{\sum_{y^n} \max_{\theta' \in \mathcal{M}} P_{\theta'}\left(y^n\right)} \tag{9}$$

to sequence x^n. This approach minimizes the individual minimax redundancy, giving individual redundancy of

$$\hat{R}_n\left(Q, x^n\right) = \frac{1}{n} \log \frac{\max_{\theta' \in \mathcal{M}} P_{\theta'}\left(x^n\right)}{Q\left(x^n\right)} = \frac{1}{n} \log \left\{ \sum_{y^n} \max_{\theta' \in \mathcal{M}} P_{\theta'}\left(y^n\right) \right\} \tag{10}$$

to every x^n, specifically achieving the individual minimax redundancy.

It is now left to bound the logarithm of the sum in (10). For the first two regions, we follow the approach used in Theorem 2 in [21] for bounding the redundancy for standard compression of i.i.d. sequences over large alphabets, but adjust it to monotonic distributions. Alternatively, one can derive the same bounds following the approach used for bounding the individual minimax redundancy of patterns in proving Theorem 12 in [20]. Let $\mathbf{n}_x^\ell \stackrel{\triangle}{=} (n_x(1), n_x(2), \ldots, n_x(\ell))$ denote the occurrence counts of the first ℓ letters of the alphabet Σ in x^n. For $\ell = k$, $\sum_{i=1}^k n_x(i) = n$.

Now, following (10),

$$\begin{aligned}
n\hat{R}_n^+\left(\mathcal{M}_k\right) &\overset{(a)}{\geq} \log\left\{\sum_{y^n:\hat{\theta}(y^n)\in\mathcal{M}} P_{\hat{\theta}}\left(y^n\right)\right\} \\
&\overset{(b)}{\geq} \log\left\{\sum_{\ell=1}^{k}\sum_{\mathbf{n}_y^\ell}\frac{1}{\ell!}\cdot\binom{n}{n_y(1),\ldots,n_y(\ell)}\cdot\prod_{i=1}^{\ell}\left(\frac{n_y(i)}{n}\right)^{n_y(i)}\right\} \\
&\overset{(c)}{\geq} \log\left\{\sum_{\mathbf{n}_y^k}\frac{1}{k!}\cdot\binom{n}{n_y(1),\ldots,n_y(k)}\cdot\prod_{i=1}^{k}\left(\frac{n_y(i)}{n}\right)^{n_y(i)}\right\} \\
&\overset{(d)}{\geq} \log\left\{\frac{1}{k!}\cdot\sum_{\mathbf{n}_y^k}\frac{\sqrt{2\pi n}}{e^{k/12}\cdot(2\pi)^{k/2}}\cdot\frac{1}{\prod_{i=1}^{k}\sqrt{n_x(i)}}\right\} \\
&\overset{(e)}{\geq} \log\left\{\frac{1}{k!}\cdot\binom{n-1}{k-1}\cdot\frac{\sqrt{2\pi n}}{e^{k/12}}\cdot\left(\frac{k}{2\pi n}\right)^{k/2}\right\} \\
&\overset{(f)}{\geq} \frac{k-1}{2}\log\frac{n}{k^3}+k\log\frac{e^{23/12}}{\sqrt{2\pi}}-O\left(\log k\right)
\end{aligned} \tag{11}$$

where (a) follows from including only sequences y^n that have a monotonic empirical (i.i.d. ML) distribution in Shtarkov's sum. Inequality (b) follows from partitioning the sequences y^n into types as done in [21], first by the number of occurring symbols ℓ, and then by the empirical distribution. Unlike standard i.i.d. distributions though, monotonicity implies that only the first ℓ symbols in Σ occur, and thus the choice of ℓ out of k in the proof in [21] is replaced by 1. Like in coding patterns, we also divide by $\ell!$ because each type with ℓ occurring symbols can be ordered in at most $\ell!$ ways, where only some retain the monotonicity. (Note that this step is the reason that step (b) produces an inequality, because more than one of the orderings may be monotonic if equal occurrence counts occur.) Except the division by $\ell!$, the remaining steps follow those in [21]. Retaining only the term $\ell = k$ yields inequality (c). Inequality (d) follows from Stirling's bound

$$\sqrt{2\pi m}\cdot\left(\frac{m}{e}\right)^m \leq m! \leq \sqrt{2\pi m}\cdot\left(\frac{m}{e}\right)^m\cdot\exp\left\{\frac{1}{12m}\right\}. \tag{12}$$

Then, (e) follows from the relation between arithmetic and geometric means, and from expressing the number of types as the number of ordered partitions of n into k parts $\binom{n-1}{k-1}$. Finally, (f) follows from applying (12) again and by lower bounding $\binom{n-1}{k-1}$.

The first region in (8) results directly from (11). The behavior is similar to patterns as shown in [1] for this region. As mentioned in [20], to obtain the second region, the bound is maximized by retaining $\hat{\ell} = \left(n^{1/3}e^{5/18}\right)/(2\pi)^{1/3}$ instead of k in step (c) of (11), for every $k \geq \hat{\ell}$. The bounds obtained are equal to those obtained for patterns because the first step (a) in (11) discards all the sequences whose contributions to Shtarkov's sum are different between patterns and monotonic distributions. A similar step is effectively done deriving the bounds for patterns. The difference is that in the case of patterns, components of Shtarkov's sum are reduced, but all are retained in the sum, while here, we omit components from the sum, corresponding to sequences with non-monotonic i.i.d. ML estimates. The analysis in [20] that also attains the second region of the bound in (8) is still valid here. It differs from the steps taken above by lower bounding a pattern probability by a larger probability than the ML i.i.d. probability corresponding to the pattern. The bound used in the derivation of Theorem 12 in [20] adds a multiplicative factor to each pattern probability which equals the number of sequences with the same pattern and an equal i.i.d. ML probability. However, this similar effect is included in Shtarkov's sum for monotonic distributions since all these sequences do have a corresponding i.i.d. ML estimate which is monotonic, and are thus not omitted by step (a) of the derivation.

The analysis in [14] yields the third region of the bound in (8), since, for $k \geq n$,

$$\begin{aligned} \hat{R}_n^+ \left(\mathcal{M}_k\right) &= \frac{1}{n} \log \left\{ \sum_{y^n} P_{\hat{\theta}_{\mathcal{M}}} \left(y^n\right) \right\} \\ &\overset{(a)}{\geq} \frac{1}{n} \log \left\{ \sum_{\Psi(y^n)} P_{\hat{\theta}} \left(y^n\right) \right\} \overset{(b)}{=} \frac{1.5 n^{1/3} \log e}{n} - O\left(\frac{\log n}{n}\right), \end{aligned} \tag{13}$$

where $\Psi(y^n)$ is the pattern of the sequence y^n. Inequality (a) holds because each pattern corresponds to at least one sequence whose ML probability parameter estimates are ordered, i.e., $\hat{\theta}_i \geq \hat{\theta}_{i+1}, \forall i$, where the most probable index represents $i = 1$, the second most probable index $i = 2$, and so on. Note that the sum element on the right hand side is for a probability of a sequence, not a pattern, but the sum is over all patterns. The left hand side also includes sequences for which the probabilities are unordered. Furthermore, exchanging the letters that correspond to two indices with the same occurrence count will not violate monotonicity. Thus the inequality follows. Step (b) in (13) is taken from [14], where the sum on the left hand side was shown to equal the right hand side. This was true when summing over all patterns with up to n indices, thus requiring $k \geq n$. Note that this requirement does not mean that n distinct symbols must occur in x^n, only that the maximal symbol in x^n is n or greater. This concludes the proof of Theorem 3. □

4 Upper Bounds for Small and Large Alphabets

In this section, we demonstrate codes that asymptotically achieve the lower bounds for $\boldsymbol{\theta} \in \mathcal{M}_k$ and $k = O(n)$. We begin with a theorem that shows the achievable redundancies, and devote the remainder of the section to describing the codes and deriving upper bounds on their redundancies. The theorem is stated assuming no initial knowledge of k. The proof first considers the setting where k is known, and then shows how the same bounds are achieved even when k is unknown in advance, but as long as it satisfies the conditions.

Theorem 4 *Fix an arbitrarily small $\varepsilon > 0$, and let $n \to \infty$. Then, there exist a code with length function $L^*(\cdot)$ that achieves redundancy*

$$R_n\left(L^*, \boldsymbol{\theta}\right) \leq \begin{cases} (1+\varepsilon)\,\frac{k-1}{2n} \log \frac{n(\log n)^2}{k^3}, & \textit{for } k \leq n^{1/3}, \\ (1+\varepsilon)\,(\log n) \left(\log \frac{k}{n^{1/3-\varepsilon}}\right) \frac{n^{1/3}}{n}, & \textit{for } n^{1/3} < k = o(n), \\ (1+\varepsilon)\,\frac{2}{3}\,(\log n)^2\, \frac{n^{1/3}}{n}, & \textit{for } n^{1/3} < k = O(n), \end{cases} \tag{14}$$

for i.i.d. sequences generated by any source $\boldsymbol{\theta} \in \mathcal{M}_k$.

Slightly tighter bounds are possible in the first and second regions and between them. The bounds presented, however, are inclusive for each of the regions. Note that the third region contains the second, but if $k = o(n)$, a tighter bound is possible in the second region. The code designed to code a sequence x^n is a two part code [23] that quantizes a distribution that minimizes the cost, and uses it to code x^n. The total redundancy cost consists of the cost of describing the quantized distribution and the quantization cost. The second is bounded through the quantized true distribution of the sequence, which cannot result in lower cost than that of the chosen distribution (which minimizes the cost). In order to achieve the low costs of the lower bound, the probability parameters are quantized non-uniformly, where the smaller the probability the finer the quantization. This approach was used in [25] and [26] to obtain upper bounds on the redundancy for coding over large alphabets and for coding patterns, respectively. The method used in [25] and [26], however, is insufficient here, because it still results in too many quantization points due to the polynomial growth in quantization spacing. Here, we use an exponential growth as the parameters increase. This general idea was used in [28] to improve an upper bound on the redundancy of coding patterns. Here, however, we improve on the method presented in [28]. Another key step in the proof here is the fact that since both encoder and decoder know the order of the probabilities *a-priori*, this order need not be coded. It is sufficient to encode the quantized probabilities of the monotonic distribution, and the decoder can identify which probability is associated with which symbol using the monotonicity of the distribution.

Proof of Theorem 4: We start with $k \leq n^{1/3}$ assuming k is known. Let $\beta = 1/(\log n)$ be a parameter (note, that we can choose other values). Partition the probability space into $J_1 = \lceil 1/\beta \rceil$ intervals,

$$I_j = \left[\frac{n^{(j-1)\beta}}{n}, \frac{n^{j\beta}}{n} \right), \quad 1 \leq j \leq J_1. \tag{15}$$

Note that $I_1 = [1/n, 2/n)$, $I_2 = [2/n, 4/n), \ldots,$ $I_j = [2^{j-1}/n, 2^j/n)$. Let $k_j = |\theta_i \in I_j|$ denote the number of probabilities in $\boldsymbol{\theta}$ that are in interval I_j. In interval j, take a grid of points with spacing

$$\Delta_j^{(1)} = \frac{\sqrt{k} n^{j\beta}}{n^{1.5}}. \tag{16}$$

Note that to complete all points in an interval, the spacing between two points at the boundary of an interval may be smaller. There are $\lceil \log n \rceil$ intervals. Ignoring negligible integer length constraints (here and elsewhere), in each interval, the number of points is bounded by

$$|I_j| \leq \frac{1}{2} \cdot \sqrt{\frac{n}{k}}, \quad \forall j : j = 1, 2, \ldots, J_1, \tag{17}$$

where $| \cdot |$ denotes the cardinality of a set. Let the *grid*

$$\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots) = \left(\frac{1}{n}, \frac{1}{n} + \frac{2\sqrt{k}}{n^{1.5}}, \ldots, \frac{2}{n}, \frac{2}{n} + \frac{4\sqrt{k}}{n^{1.5}}, \ldots \right) \tag{18}$$

be a vector that takes all the points from all intervals, with cardinality

$$B_1 \stackrel{\triangle}{=} |\boldsymbol{\tau}| \leq \frac{1}{2} \cdot \sqrt{\frac{n}{k}} \lceil \log n \rceil. \tag{19}$$

Now, let $\boldsymbol{\varphi} = (\varphi_1, \varphi_2, \ldots, \varphi_k)$ be a monotonic probability vector, such that $\sum \varphi_i = 1$, $\varphi_1 \geq \varphi_2 \geq \cdots \geq \varphi_k \geq 0$, and also the smaller $k-1$ components of $\boldsymbol{\varphi}$ are either 0 or from $\boldsymbol{\tau}$, i.e., $\varphi_i \in (\boldsymbol{\tau} \cup \{0\})$, $i = 2, 3, \ldots, k$. One can code x^n using a two part code, assuming the distribution governing x^n is given by the parameter $\boldsymbol{\varphi}$. The code length required (up to integer length constraints) is

$$L(x^n | \boldsymbol{\varphi}) = \log k + L_R(\boldsymbol{\varphi}) - \log P_{\varphi}(x^n), \tag{20}$$

where $\log k$ bits are needed to describe how many letter probabilities are greater than 0 in $\boldsymbol{\varphi}$, and $L_R(\boldsymbol{\varphi})$ is the number of bits required to describe the quantized points of $\boldsymbol{\varphi}$.

The vector $\boldsymbol{\varphi}$ can be described by a code as follows. Let $\hat{k}_{\varphi}$ be the number of nonzero letter probabilities hypothesized by $\boldsymbol{\varphi}$. Let b_i denote the index of φ_i in $\boldsymbol{\tau}$, i.e., $\varphi_i = \tau_{b_i}$. Then, we will use the following differential code. For $\varphi_{\hat{k}_{\varphi}}$ we need at most $1 + \log b_{\hat{k}_{\varphi}} + 2\log(1 + \log b_{\hat{k}_{\varphi}})$

bits to code its index in $\boldsymbol{\tau}$ using Elias' coding for the integers [7]. For φ_{i-1}, we need at most $1 + \log(b_{i-1} - b_i + 1) + 2\log[1 + \log(b_{i-1} - b_i + 1)]$ bits to code the index displacement from the index of the previous parameter, where an additional 1 is added to the difference in case the two parameters share the same index. Summing up all components of $\boldsymbol{\varphi}$, and taking $b_{\hat{k}_\varphi+1} = 0$,

$$\begin{aligned} L_R(\boldsymbol{\varphi}) &\leq \hat{k}_\varphi - 1 + \sum_{i=2}^{\hat{k}_\varphi} \log\left(b_i - b_{i+1} + 1\right) + 2\sum_{i=2}^{\hat{k}_\varphi} \log\left[1 + \log\left(b_i - b_{i+1} + 1\right)\right] \\ &\overset{(a)}{\leq} (k-1) + (k-1)\log\frac{B_1 + k - 1}{k} + 2(k-1)\log\log\frac{B_1 + k - 1}{k} + o(k) \\ &\overset{(b)}{=} (1+\varepsilon)\frac{k-1}{2}\log\frac{n\left(\log n\right)^2}{k^3}. \end{aligned} \tag{21}$$

Inequality (a) is obtained by applying Jensen's inequality once on the first sum, twice on the second sum utilizing the monotonicity of the logarithm function, and by bounding $\hat{k}_\varphi$ by k and absorbing low order terms in the resulting $o(k)$ term. Then, low order terms are absorbed in ε, and (19) is used to obtain (b).

To code x^n, we choose $\boldsymbol{\varphi}$ which minimizes the expression in (20) over all $\boldsymbol{\varphi}$, i.e.,

$$L^*\left(x^n\right) = \min_{\boldsymbol{\varphi}} L\left(x^n|\boldsymbol{\varphi}\right) \overset{\triangle}{=} L\left(x^n|\hat{\boldsymbol{\varphi}}\right). \tag{22}$$

The *pointwise* redundancy for x^n is given by

$$nR_n\left(L^*, x^n\right) = L^*\left(x^n\right) + \log P_\theta\left(x^n\right) = \log k + L_R^*\left(\hat{\boldsymbol{\varphi}}\right) + \log\frac{P_\theta\left(x^n\right)}{P_{\hat{\varphi}}\left(x^n\right)}. \tag{23}$$

Note that the pointwise redundancy differs from the individual one, since it is defined w.r.t. the true probability of x^n.

To bound the third term of (23), let $\boldsymbol{\theta}'$ be a quantized still monotonic version of $\boldsymbol{\theta}$ onto $\boldsymbol{\tau}$, i.e., $\theta'_i \in (\boldsymbol{\tau} \cup \{0\}),\ i = 2, 3, \ldots, k$, where if $\theta_i > 0 \Leftrightarrow \theta'_i > 0$ as well. Define the quantization error,

$$\delta_i = \theta_i - \theta'_i. \tag{24}$$

The quantization is performed from the smallest parameter θ_k to the largest, where monotonicity is retained, as well as minimal absolute quantization error. This implies that θ_i will be quantized to one of the two nearest grid points (one smaller and one greater than it). It also guarantees that $|\delta_1| \leq \Delta_{j_2}^{(1)}$, where j_2 is the index of the interval in which θ_2 is contained, i.e., $\theta_2 \in I_{j_2}$. Now, since $\boldsymbol{\theta}'$ is included in the minimization of (22), we have, for every x^n,

$$L^*\left(x^n\right) \leq L\left(x^n|\boldsymbol{\theta}'\right), \tag{25}$$

and also

$$nR_n\left(L^*, x^n\right) \leq \log k + L_R\left(\boldsymbol{\theta}'\right) + \log \frac{P_\theta\left(x^n\right)}{P_{\theta'}\left(x^n\right)}. \tag{26}$$

Averaging over all possible x^n, the average redundancy is bounded by

$$\begin{aligned} nR_n\left(L^*, \boldsymbol{\theta}\right) &= \log k + E_\theta L_R^*\left(\hat{\boldsymbol{\varphi}}\right) + E_\theta \log \frac{P_\theta\left(X^n\right)}{P_{\hat{\varphi}}\left(X^n\right)} \\ &\leq \log k + E_\theta L_R\left(\boldsymbol{\theta}'\right) + E_\theta \log \frac{P_\theta\left(X^n\right)}{P_{\theta'}\left(X^n\right)}. \end{aligned} \tag{27}$$

The second term of (27) is bounded with the bound of (21), and we proceed with the third term.

$$\begin{aligned} E_\theta \log \frac{P_\theta\left(X^n\right)}{P_{\theta'}\left(X^n\right)} &\overset{(a)}{=} n\sum_{i=1}^{k} \theta_i \log \frac{\theta_i}{\theta_i'} \overset{(b)}{=} n\sum_{i=1}^{k}\left(\theta_i' + \delta_i\right)\log\left(1 + \frac{\delta_i}{\theta_i'}\right) \\ &\overset{(c)}{\leq} n(\log e)\sum_{i=1}^{k}\left(\theta_i' + \delta_i\right)\frac{\delta_i}{\theta_i'} \overset{(d)}{=} n(\log e)\sum_{i=1}^{k}\frac{\delta_i^2}{\theta_i'} \\ &\overset{(e)}{\leq} k\log e + \frac{2(\log e)k}{n}\sum_{j=1}^{J_1} k_j \cdot n^{j\beta} \overset{(f)}{\leq} 5(\log e)k. \end{aligned} \tag{28}$$

Equality (a) is since the argument in the logarithm is fixed, thus expectation is performed only on the number of occurrences of letter i for each letter. Representing $\theta_i = \theta_i' + \delta_i$ yields equation (b). We use $\ln(1+x) \leq x$ to obtain (c). Equality (d) is obtained since all the quantization displacements must sum to 0. The first term of inequality (e) is obtained under a worst case assumption that $\theta_i \ll 1/n$ for $i \geq 2$. Thus it is quantized to $\theta_i' = 1/n$, and the bound $|\delta_i| \leq 1/n$ is used. The second term is obtained by separating the terms into their intervals. In interval j, the bounds $\theta_i' \geq n^{(j-1)\beta}/n$, and $|\delta_i| \leq \sqrt{k}n^{j\beta}/n^{1.5}$ are used, and also $n^\beta = 2$. Inequality (f) is obtained since

$$\sum_{j=1}^{J_1} k_j n^{j\beta} = \sum_{j=1}^{J_1} k_j 2^j \leq 2n. \tag{29}$$

Inequality (29) is obtained since $k_1 \leq n$, $k_2 \leq (n-k_1)/2$, $k_3 \leq (n-k_1)/4 - k_2/2$, and so on, until

$$k_{J_1} \leq \frac{n}{2^{J_1-1}} - \sum_{\ell=1}^{J_1}\frac{k_\ell}{2^{J_1-\ell}} \Rightarrow \sum_{j=1}^{J_1} k_j 2^j \leq 2n. \tag{30}$$

The reason for these relations are the lower limits of the J_1 intervals that restrict the number of parameters inside the interval. The restriction is done in order of intervals, so that the used probabilities are subtracted, leading to the series of equations.

Plugging the bounds of (21) and (28) into (27), we obtain,

$$\begin{aligned} nR_n\left(L^*, \boldsymbol{\theta}\right) &\leq \log k + (1+\varepsilon)\frac{k-1}{2}\log\frac{n\left(\log n\right)^2}{k^3} + 5(\log e)k \\ &\leq \left(1+\varepsilon'\right)\frac{k-1}{2}\log\frac{n\left(\log n\right)^2}{k^3}, \end{aligned} \tag{31}$$

where we absorb low order terms in ε'. Replacing ε' by ε normalizing the redundancy per symbol by n, the bound of the first region of (14) is proved.

We now consider the larger values of k, i.e., $n^{1/3} < k = O(n)$. The idea of the proof is the same. However, we need to partition the probability space to different intervals, the spacing within an interval must be optimized, and the parameters' description cost must be bounded differently, because now there are more parameters quantized than points in the quantization grid. Define the jth interval as

$$I_j = \left[\frac{n^{(j-1)\beta}}{n^2}, \frac{n^{j\beta}}{n^2}\right), \quad 1 \leq j \leq J_2, \tag{32}$$

where $J_2 = \lceil 2/\beta \rceil = \lceil 2 \log n \rceil$. Again, let $k_j = |\theta_i \in I_j|$ denote the number of probabilities in $\boldsymbol{\theta}$ that are in interval I_j. It could be possible to use the intervals as defined in (15), but this would not guarantee bounded redundancy in the rate we require if there are very small probabilities $\theta_i \ll 1/n$. Therefore, the interval definition in (15) can be used for larger alphabets only if the probabilities of the symbols are known to be bounded. Define the spacing in interval j as

$$\Delta_j^{(2)} = \frac{n^{j\beta}}{n^{2+\alpha}}, \tag{33}$$

where α is a parameter to be optimized. Similarly to (17), the interval cardinality here is

$$|I_j| \leq 0.5 \cdot n^{\alpha}, \quad \forall j : j = 1, 2, \ldots, J_2, \tag{34}$$

In a similar manner to the definition of $\boldsymbol{\tau}$ in (18), we define

$$\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots) = \left(\frac{1}{n^2}, \frac{1}{n^2} + \frac{2}{n^{2+\alpha}}, \ldots, \frac{2}{n^2}, \frac{2}{n^2} + \frac{4}{n^{2+\alpha}}, \ldots\right). \tag{35}$$

The cardinality of $\boldsymbol{\eta}$ is

$$B_2 \stackrel{\triangle}{=} |\boldsymbol{\eta}| \leq 0.5 \cdot n^{\alpha} \lceil 2 \log n \rceil \leq n^{\alpha} \lceil \log n \rceil. \tag{36}$$

We now perform the encoding similarly to the small k case, where we allow quantization to nonzero values to the components of $\boldsymbol{\varphi}$ up to $i = n^2$. (This is more than needed but is possible since $\eta_1 = 1/n^2$.) Encoding is performed similarly to the small k case. Thus, similarly to (27), we have

$$nR_n(L^*, \boldsymbol{\theta}) \leq 2 \log n + E_\theta L_R(\boldsymbol{\theta}') + E_\theta \log \frac{P_\theta(X^n)}{P_{\theta'}(X^n)}, \tag{37}$$

where the first term is due to allowing up to $\hat{k} = n^2$. Since usually in this region $k \geq B_2$ (except the low end), the description of vectors $\boldsymbol{\varphi}$ and $\boldsymbol{\theta}'$ is done by coding the cardinality of $|\varphi_i = \eta_j|$ and $|\theta'_i = \eta_j|$, respectively, i.e., for each grid point the code describes how many letters have probability quantized to this point. This idea resembles coding profiles of patterns, as done in [20]. However, unlike the method in [20], here, many probability parameters of symbols with different occurrences are mapped to the same grid point by quantization. The number of parameters mapped to a grid point of $\boldsymbol{\eta}$ is coded using Elias' representation of the integers. Hence, in a similar manner to (21),

$$\begin{aligned}
L_R(\boldsymbol{\theta}') &\overset{(a)}{\leq} \sum_{j=1}^{B_2} \left\{1 + \log\left(|\theta'_i = \eta_j| + 1\right) + 2\log\left[1 + \log\left(|\theta'_i = \eta_j| + 1\right)\right]\right\} \\
&\overset{(b)}{\leq} B_2 + B_2 \log \frac{k + B_2}{B_2} + 2B_2 \log\log \frac{k + B_2}{B_2} + o\left(B_2\right) \\
&\overset{(c)}{\leq} \begin{cases} (1+\varepsilon)(\log n)\left(\log \frac{k}{n^{\alpha-\varepsilon}}\right) n^{\alpha}, & \text{for } n^{\alpha} < k = o(n), \\ (1+\varepsilon)(1-\alpha)\left(\log n\right)^2 n^{\alpha}, & \text{for } n^{\alpha} < k = O(n). \end{cases}
\end{aligned} \tag{38}$$

The additional 1 term in the logarithm in (a) is for 0 occurrences, (b) is obtained similarly to step (a) of (21), absorbing all low order terms in the last term. To obtain (c), we first assume, for the first region, that $kn^{\varepsilon} \gg B_2$ (an assumption that must be later validated with the choice of α). Then, low order terms are absorbed in ε. The extra n^{ε} factor is unnecessary if $k \gg B_2$. The second region is obtained by upper bounding k without this factor. It is possible to separate the first region into two regions, eliminate this factor in the lower region, and obtain a more complicated, yet tighter, expression in the upper region, where $k \sim \Theta(n^{1/3})$.

Now, similarly to (28), we obtain

$$\begin{aligned}
E_{\theta} \log \frac{P_{\theta}\left(X^n\right)}{P_{\theta'}\left(X^n\right)} &\leq n(\log e) \sum_{i=1}^{k} \frac{\delta_i^2}{\theta'_i} \\
&\overset{(a)}{\leq} O(1) + \frac{2\log e}{n^{1+2\alpha}} \sum_{j=1}^{J_2} k_j n^{j\beta} \overset{(b)}{\leq} 4(\log e) n^{1-2\alpha} + O(1).
\end{aligned} \tag{39}$$

The first term of inequality (a) is obtained under the assumption that $k = O(n)$, $\theta'_i \geq 1/n^2$, and $|\delta_i| \leq 1/n^2$. For the second term $|\delta_i| \leq n^{j\beta}/n^{2+\alpha}$, and $\theta'_i \geq n^{(j-1)\beta}/n^2$. Inequality (b) is obtained in a similar manner to inequality (f) of (28), where the sum is shown similarly to be $2n^2$.

Summing up the contributions of (38) and (39) in (37), it is clear that $\alpha = 1/3$ minimizes the total cost (to first order). This choice of α also satisfies the assumption of step (c) in (38). Using $\alpha = 1/3$, absorbing all low order terms in ε and normalizing by n, we obtain the remaining two regions of the bound in (14). It should be noted that the proof here would give a bound of $O(n^{1/3+\varepsilon})$ up to $k = O(n^{4/3})$. If the intervals in (15) were used for bounded distributions, the coefficients of the last two regions will be reduced by a factor of 2. Additional manipulations on the grid $\boldsymbol{\eta}$ may reduce the coefficients more (see, e.g., [28]).

The proof up to this point assumes that k is known in advance. This is important for the code resulting in the bound for the first region because the quantization grid depends on k. Specifically, if in building the grid, k is underestimated, the description cost of $\boldsymbol{\varphi}$ increases. If k is overestimated, the quantization cost will increase. Also, if the code of the second region is used for a smaller k, a larger bound than necessary results. To solve this, the optimization that chooses $L^*(x^n)$ is done over all possible values of k (greater than or equal to the maximal symbol occurring in x^n), i.e., every greater k in the first region, and the construction of the code for the other regions. For every k in the first region, a different construction is done, using the appropriate k to determine the spacing in each interval. The value of k yielding the shortest code word is then used, and $O(\log n)$ additional bits are used at the prefix of the code to inform the decoder which k is used. The analysis continues as before. This does not change the redundancy to first order, giving all three regions of the bound in (14), even if k is unknown in advance. This concludes the proof of Theorem 4. □

5 Upper Bounds for Fast Decaying Distributions

This section shows that with some mild conditions on the source distribution, the same redundancy upper bounds achieved for finite monotonic distributions can be achieved even if the monotonic distribution is over an infinite alphabet. The key observation that allows this is that a distribution that decays fast enough will result in only a small number of occurrences of unlikely letters in a sequence. These letters may very likely be out of order, but since there are very few of them, they can be handled without increasing the asymptotic behavior of the coding cost. More precisely, fast decaying monotonic distributions can be viewed as if they have some effective bounded alphabet size, where occurrences of symbols outside this limited alphabet are rare. We present two theorems and a corollary that show how one can upper bound the redundancy obtained when coding with some unknown distribution. The first theorem provides a slightly stronger bound (with smaller coefficient) even for $k = O(n)$, where the smaller coefficient is attained by improved bounding, that more uniformly weights the quantization cost for minimal probabilities. In the weaker version of the results presented here, if the distribution decays slower and there are more low probability symbols, the redundancy order does increase due to the penalty of identifying these symbols in a sequence. However, we show, consistently with the results in [10], that as long as the entropy of the source is finite, a universal code, in the sense of diminishing redundancy per symbol, does exist. We begin with stating the two theorems and the corollary, then the proofs are presented.

The section is concluded with three examples of typical monotonic distributions over the integers, to which the bounds are applied.

5.1 Upper Bounds

We begin with some notation. Fix an arbitrary small $\varepsilon > 0$, and let $n \to \infty$. Define $m \stackrel{\triangle}{=} m_\rho \stackrel{\triangle}{=} n^\rho$ as the *effective alphabet size*, where $\rho > \varepsilon$. (Note that $\rho = (\log m)/(\log n)$.) Let

$$\mathcal{R}_n(m) \stackrel{\triangle}{=} \begin{cases} \frac{m-1}{2} \log \frac{n}{m^3}, & \text{for } m = o\left(n^{1/3}\right), \\ \frac{1}{2} \cdot \left(\rho + \frac{2}{3}\right)\left(\rho + \varepsilon - \frac{1}{3}\right)(\log n)^2 n^{1/3}, & \text{otherwise.} \end{cases} \tag{40}$$

Theorem 5 *I. Fix an arbitrarily small $\varepsilon > 0$, and let $n \to \infty$. Let x^n be generated by an i.i.d. monotonic distribution $\boldsymbol{\theta} \in \mathcal{M}$. If there exists m^*, such that,*

$$\sum_{i > m^*} n\theta_i \log i = o\left[\mathcal{R}_n\left(m^*\right)\right], \tag{41}$$

then, there exists a code with length function $L^*(\cdot)$, such that

$$R_n\left(L^*, \boldsymbol{\theta}\right) \leq \frac{(1+\varepsilon)}{n} \mathcal{R}_n\left(m^*\right) \tag{42}$$

for the monotonic distribution $\boldsymbol{\theta}$.

II. If there exists m^* for which $\rho^* = o\left(n^{1/3}/(\log n)\right)$, such that,

$$\sum_{i > m^*} \theta_i \log i = o(1), \tag{43}$$

then, there exists a universal code with length function $L^*(\cdot)$, such that

$$R_n\left(L^*, \boldsymbol{\theta}\right) = o(1). \tag{44}$$

Theorem 5 implies that if a monotonic distribution decays fast enough, its effective alphabet size does not exceed $O(n^\rho)$, and, as long as ρ is fixed, bounds of the same order as those obtained for finite alphabets are achievable. Specifically, very fast decaying distributions, although over infinite alphabets, may even behave like monotonic distributions with $o\left(n^{1/3}\right)$ symbols. The condition in (41) merely means that the cost that a code would obtain in order to code very rare symbols, that are larger than the effective alphabet size, is negligible w.r.t. the total cost obtained from other, more likely, symbols. Note that for $m = n$, the bound is tighter than that of the third region of Theorem 4, and a constant of 5/9 replaces 2/3. The second part of the theorem states that if the decay is slow, but the cost of coding rare symbols is still diminishing per symbol, a universal code still exists for such distributions. However, in this case the redundancy will be dominated by coding the rare (out of order) symbols. This result leads to the following corollary:

Corollary 1 *As $n \to \infty$, sequences generated by monotonic distributions with $H_\theta(X) = O(1)$ are universally compressible in the average sense.*

Corollary 1 shows that sequences generated by finite entropy monotonic distributions can be compressed in the average with diminishing per symbol redundancy. This result is consistent with the results shown in [10].

While Theorem 5 bounds the redundancy decay rate with two extremes, a more general theorem can be used to provide some best redundancy decay rate that a code can be designed to adapt to for some unknown monotonic distribution that governs the data. As the examples at the end of this section show, the next theorem is very useful for slower decaying distributions.

Theorem 6 *Fix an arbitrarily small $\varepsilon > 0$, and let $n \to \infty$. Let x^n be generated by an i.i.d. monotonic distribution $\boldsymbol{\theta} \in \mathcal{M}$. Then, there exists a code with length function $L^*(\cdot)$, that achieves redundancy*

$$nR_n\left(L^*, \boldsymbol{\theta}\right) \leq (1+\varepsilon) \cdot \min_{\alpha,\rho:\rho\geq\alpha+\varepsilon} \left\{ \frac{1}{2} \cdot (\rho + 2\alpha)\,(\rho - \alpha)\,(\log n)^2 n^\alpha + 5(\log e) n^{1-2\alpha} + \left(1 + \frac{1}{\rho}\right) n \sum_{i > n^\rho} \theta_i \log i \right\} \quad (45)$$

for coding sequences generated by the source $\boldsymbol{\theta}$.

We continue with proving the two theorems and the corollary.

Proof : The idea of the proof of both theorems is to separate the more likely symbols from the unlikely ones. First, the code determines the point of separation $m = n^\rho$. (Note that ρ can be greater than 1.) Then, all symbols $i \leq m$ are considered likely and are quantized in a similar manner as in the codes for smaller alphabets. Unlike bounded alphabets, though, a more robust grid is used here to allow larger values of m. Coding of occurrences of these symbols uses the quantized probabilities. The unlikely symbols are coded hierarchically. They are first merged into a single symbol, and then are coded within this symbol, where the full cost of conveying to the decoder which rare symbols occur in the sequence is required. Thus, they are presented giving their actual value. As long as the decay is fast enough, the average cost of conveying these symbols becomes negligible w.r.t. the cost of coding the likely symbols. If the decay is slower, but still fast enough, as the case described in condition (43), the coding cost of the rare symbols dominates the redundancy, but still diminishing redundancy can be achieved. In order to determine the best value of m for a given sequence, all values are tried and the one yielding the shortest description is used for coding the specific sequence x^n.

Let $m \geq 2$ determine the number of likely symbols in the alphabet. For a given m, define

$$S_m \stackrel{\triangle}{=} \sum_{i>m} \theta_i, \tag{46}$$

as the total probability of the remaining symbols. Given $\boldsymbol{\theta}$, m and S_m, a probability

$$P\left(x^n | m, S_m, \boldsymbol{\theta}\right) \stackrel{\triangle}{=} \left[\prod_{i=1}^{m} \theta_i^{n_x(i)}\right] \cdot S_m^{n_x(x>m)} \cdot \prod_{i>m} \left(\frac{n_x(i)}{n_x(x>m)}\right)^{n_x(i)}, \tag{47}$$

can be computed for x^n, where $n_x(i)$ is the occurrence count of symbol i in x^n, and $n_x(x > m)$ is the count of all symbols greater than m in x^n. This probability mass function clusters all large symbols (with small probabilities) greater than m into one symbol. Then, it uses the ML estimate of each of the large symbols to distinguish among them in the clustered symbol.

For every m, we can define a quantization grid $\boldsymbol{\xi}_m$ for the first m probability parameters of $\boldsymbol{\theta}$. The idea is similar to that used for all probability parameters in the proof of Theorem 4. If $m = o(n^{1/3})$, we use $\boldsymbol{\xi}_m = \boldsymbol{\tau}_m$, where $\boldsymbol{\tau}_m$ is the grid defined in (18) where m replaces k. Otherwise, we can use the definition of $\boldsymbol{\eta}$ in (35). However, to obtain tighter bounds for large m, we define a different grid for the larger values of m following similar steps to those in (32)-(36). First, define the jth interval as

$$I_j = \left[\frac{n^{(j-1)\beta}}{n^{\rho+2\alpha}}, \frac{n^{j\beta}}{n^{\rho+2\alpha}}\right), \quad 1 \leq j \leq J_\rho, \tag{48}$$

where $\rho = (\log m)/(\log n)$ as defined above, α is a parameter, and $\beta = 1/(\log n)$ as before. Within the jth interval, we define the spacing in the grid by

$$\Delta_j^{(\rho)} = \frac{n^{j\beta}}{n^{\rho+3\alpha}}. \tag{49}$$

As in (34),

$$|I_j| \leq 0.5 \cdot n^\alpha, \quad \forall j : j = 1, 2, \ldots, J_\rho, \tag{50}$$

and the total number of intervals is

$$J_\rho = \lceil (\rho + 2\alpha) \log n \rceil. \tag{51}$$

Similarly to (35), $\boldsymbol{\xi}_m$ is defined as

$$\boldsymbol{\xi}_m = (\xi_1, \xi_2, \ldots) = \left(\frac{1}{n^{\rho+2\alpha}}, \frac{1}{n^{\rho+2\alpha}} + \frac{2}{n^{\rho+3\alpha}}, \ldots, \frac{2}{n^{\rho+2\alpha}}, \frac{2}{n^{\rho+2\alpha}} + \frac{4}{n^{\rho+3\alpha}}, \ldots\right). \tag{52}$$

The cardinality of $\boldsymbol{\xi}_m$ is thus

$$B_\rho \stackrel{\triangle}{=} |\boldsymbol{\xi}_m| \leq 0.5 \cdot n^\alpha \lceil (\rho + 2\alpha) \log n \rceil. \tag{53}$$

An mth order quantized version $\boldsymbol{\theta}'_m$ of $\boldsymbol{\theta}$ is obtained by quantizing θ_i, $i = 2, 3, \ldots, m$ onto $\boldsymbol{\xi}_m$, such that $\theta'_i \in \boldsymbol{\xi}_m$ for these values of i. Then, the remaining cluster probability S_m is quantized into $S'_m \in [1/n, 2/n, \ldots, 1]$. The parameter θ'_1 is constrained by the quantization of the other parameters. Quantization is performed in a similar manner as before, to minimize the accumulating cost and retain monotonicity.

Now, for any $m \geq 2$, let $\boldsymbol{\varphi}_m$ be any monotonic probability vector of cardinality m whose last $m-1$ components are quantized into $\boldsymbol{\xi}_m$, and let $\sigma_m \in [1/n, 2/n, \ldots, 1]$ be a quantized estimate of the total probability of the remaining symbols, such that $\sum_{i=1}^{m} \varphi_{i,m} + \sigma_m = 1$, where $\varphi_{i,m}$ is the ith component of $\boldsymbol{\varphi}_m$. If m, σ_m and $\boldsymbol{\varphi}_m$ are known, a given x^n can be coded using $P\left(x^n | m, \sigma_m, \boldsymbol{\varphi}_m\right)$ as defined in (47), where σ_m replaces S_m, and the m components of $\boldsymbol{\varphi}_m$ replace the first m components of $\boldsymbol{\theta}$. However, in the universal setting, none of these parameter are known in advance. Furthermore, neither the symbols greater than m nor their conditional ML probabilities are known in advance. Therefore, the total cost of coding x^n using these parameters requires universality costs for describing them. The cost of universally coding x^n assigning probability $P\left(x^n | m, \sigma_m, \boldsymbol{\varphi}_m\right)$ to it thus requires the following five components: 1) m should be described using Elias' representation with at most $1 + \rho \log n + 2 \log(1 + \rho \log n)$ bits. 2) The value of σ_m in its quantization grid should be coded using $\log n$ bits. 3) The m components of $\boldsymbol{\varphi}_m$ require $L_R\left(\boldsymbol{\varphi}_m\right)$ (which is bounded below) bits. 4) The number $c_x(x > m)$ of distinct letters in x^n greater than m is coded using $\log n$ bits. 5) Each letter $i > m$ in x^n is coded. Elias' coding for the integers using $1 + \log i + 2 \log(1 + \log i)$ bits can be used, but to simplify the derivation we can also use the code, also presented in [7], that uses no more than $1 + 2 \log i$ bits to describe i. In addition, at most $\log n$ bits are required for describing $n_x(i)$ in x^n. For $n \to \infty$, $m \gg 1$, and $\varepsilon > 0$ arbitrarily small, this yields a total cost of

$$\begin{aligned} L\left(x^n | m, \sigma_m, \boldsymbol{\varphi}_m\right) \quad \leq \quad & -\log P\left(x^n | m, \sigma_m, \boldsymbol{\varphi}_m\right) + L_R\left(\boldsymbol{\varphi}_m\right) + \left[(1+\varepsilon)\rho + c_x(x > m) + 2\right] \log n \\ & + c_x(x > m) + 2 \sum_{i > m, i \in x^n} \log i, \end{aligned} \tag{54}$$

where we assume m is large enough to bound the cost of describing m by $(1+\varepsilon)\rho \log n$.

The description cost of $\boldsymbol{\varphi}_m$ for $m = o(n^{1/3})$ is bounded by

$$L_R\left(\boldsymbol{\varphi}_m\right) \leq (1+\varepsilon) \frac{m-1}{2} \log \frac{n}{m^3} \tag{55}$$

using (21), where m replaces k. The $(\log n)^2$ factor in (21) can be absorbed in ε since we limit m to $o(n^{1/3})$, unlike the derivation in (21). For larger values of m, we describe symbol probabilities of $\boldsymbol{\varphi}_m$ in the grid $\boldsymbol{\xi}_m$ in a similar manner to the description of $O(n)$ symbol probabilities in the grid

$\boldsymbol{\eta}$. Similarly to (38), we thus have

$$\begin{aligned} L_R(\boldsymbol{\varphi}_m) &\leq B_\rho + B_\rho \log \frac{n^\rho + B_\rho}{B_\rho} + 2B_\rho \log\log \frac{n^\rho + B_\rho}{B_\rho} + o\left(B_\rho\right) \\ &\overset{(a)}{\leq} \frac{(1+\varepsilon)}{2}\left(\rho + 2\alpha\right)\left(\rho + \varepsilon - \alpha\right)(\log n)^2 n^\alpha \end{aligned} \tag{56}$$

where to obtain inequality (a), we first multiply n^ρ by n^ε in the numerator of the argument of the logarithm. This is only necessary for $\rho \to \alpha$ to guarantee that $n^{\rho+\varepsilon} \gg B_\rho$. Substituting the bound on B_ρ from (53), absorbing low order terms in the leading ε, yields the bound.

A sequence x^n can now be coded using the universal parameters that minimize the length of the sequence description, i.e.,

$$L^*\left(x^n\right) \overset{\triangle}{=} \min_{m' \geq 2} \min_{\sigma_{m'} \in \left[\frac{1}{n}, \frac{2}{n}, \ldots, 1\right]} \min_{\boldsymbol{\varphi}_{m'} : \varphi_i \in \boldsymbol{\xi}_{m'}, i \geq 2} L\left(x^n | m', \sigma_{m'}, \boldsymbol{\varphi}_{m'}\right) \leq L\left(x^n | m, S'_m, \boldsymbol{\theta}'_m\right), \tag{57}$$

where $\boldsymbol{\theta}'_m$ and S'_m are the true source parameters quantized as described above, and the inequality holds for every m. Note that the maximization on m' should be performed only up to the maximal symbol the occurs in x^n.

Following (54)-(57), up to negligible integer length constraints, the average redundancy using $L^*(\cdot)$ is bounded, for every $m \geq 2$, by

$$\begin{aligned} nR_n\left(L^*, \boldsymbol{\theta}\right) &= E_\theta\left[L^*\left(X^n\right) + \log P_\theta\left(X^n\right)\right] \\ &\overset{(a)}{\leq} E_\theta\left[L\left(X^n \mid m, S'_m, \boldsymbol{\theta}'_m\right) + \log P_\theta\left(X^n\right)\right] \\ &\overset{(b)}{\leq} E_\theta \log \frac{P_\theta\left(X^n\right)}{P\left(X^n \mid m, S'_m, \boldsymbol{\theta}'_m\right)} + L_R\left(\boldsymbol{\theta}'_m\right) + 2\sum_{i>m} P_\theta\left(i \in X^n\right) \log i \\ &\quad + (1+\varepsilon)\left[E_\theta C_x\left(X > m\right) + \rho + 2\right] \log n \end{aligned} \tag{58}$$

where (a) follows from (57), and (b) follows from averaging on (54) with $\sigma_m = S'_m$, and $\boldsymbol{\varphi}_m = \boldsymbol{\theta}'_m$, where the average on $c_x(x > m)$ is absorbed in the leading ε.

Expressing $P_\theta\left(x^n\right)$ as

$$P_\theta\left(x^n\right) = \left[\prod_{i \leq m} \theta_i^{n_x(i)}\right] \cdot S_m^{n_x(x>m)} \cdot \prod_{i>m} \left(\frac{\theta_i}{S_m}\right)^{n_x(i)}, \tag{59}$$

and defining $\delta_S \stackrel{\triangle}{=} S_m - S'_m$, the first term of (58) is bounded, for the upper region of m, by

$$\begin{aligned}
E_\theta \log \frac{P_\theta\left(X^n\right)}{P\left(X^n \mid m, S'_m, \boldsymbol{\theta}'_m\right)} &\leq E_\theta \left[\sum_{i=1}^{m} N_x(i) \log \frac{\theta_i}{\theta'_{i,m}} + N_x\left(X > m\right) \log \frac{S_m}{S'_m} + \right. \\
&\left. \sum_{i>m} N_x(i) \log \frac{\theta_i / S_m}{N_x(i)/N_x(X>m)} \right] \\
&\stackrel{(a)}{\leq} n \cdot \sum_{i=1}^{m} \theta_i \log \frac{\theta_i}{\theta'_{i,m}} + n S_m \log \frac{S_m}{S'_m} \\
&\stackrel{(b)}{\leq} n(\log e) \left[\left(\sum_{i=1}^{m} \frac{\delta_i^2}{\theta'_{i,m}} \right) + \frac{\delta_S^2}{S'_m} \right] \\
&\stackrel{(c)}{\leq} (\log e) \cdot \frac{n \cdot n^\rho}{n^{\rho+2\alpha}} + 2(\log e) n^{1-\rho-4\alpha} \cdot \sum_{j=1}^{J_\rho} k_j n^{j\beta} + \log e \\
&\stackrel{(d)}{\leq} 5(\log e) n^{1-2\alpha} + \log e, \qquad (60)
\end{aligned}$$

where (a) is since for the third term, the conditional ML probability used for coding is greater than the actual conditional probability assigned to all letters greater than m for every x^n. Hence, the third term is bounded by 0. For the other terms expectation is performed. Inequality (b) is obtained similarly to (28) where quantization includes the first m components of $\boldsymbol{\theta}$ and the parameter S_m. Then, inequality (c) follows the same reasoning as step (a) of (39). The first term bounds the worst case in which all n^ρ symbols are quantized to $1/n^{\rho+2\alpha}$ with $|\delta_i| \leq 1/n^{\rho+2\alpha}$. The second term is obtained where $\theta'_{i,m} \geq n^{(j-1)\beta}/n^{\rho+2\alpha}$ and $|\delta_i| \leq n^{j\beta}/n^{\rho+3\alpha}$ for $\theta_i \in I_j$, and $k_j = |\theta_i \in I_j|$ as before. The last term is since $S'_m \geq 1/n$ and $|\delta_S| \leq 1/n$. Finally, (d) is obtained similarly to step (b) of (39), where as in (29), $\sum k_j n^{j\beta} \leq 2n^{\rho+2\alpha}$. For $m = o(n^{1/3})$, the same initial steps up to step (b) in (60) are applied, and then the remaining steps in (28) are applied to the left sum with m replacing k, yielding a total quantization cost of $5(\log e)m + \log e$.

To bound the third and fourth terms of (58), we realize that

$$P_\theta\left(i \in X^n\right) = 1 - \left(1 - \theta_i\right)^n \leq n\theta_i. \qquad (61)$$

Similarly,

$$E_\theta C_x(X > m) = \sum_{i>m} P_\theta\left(i \in X^n\right) \leq n S_m. \qquad (62)$$

Combining the dominant terms of the third and fourth terms of (58), we have

$$\begin{aligned}
& 2\sum_{i>m} P_\theta\left(i \in X^n\right)\log i + (1+\varepsilon)E_\theta C_x(X>m)\log n \\
& \overset{(a)}{=} \quad \sum_{i>m} P_\theta\left(i \in X^n\right)\left[2\log i + (1+\varepsilon)\log n\right] \\
& \overset{(b)}{\leq} \quad \left(2+\frac{1+\varepsilon}{\rho}\right)\sum_{i>m} P_\theta\left(i \in X^n\right)\log i \overset{(c)}{\leq} \left(2+\frac{1+\varepsilon}{\rho}\right) n \sum_{i>m}\theta_i \log i
\end{aligned} \tag{63}$$

where (a) is because $E_\theta C_x(X > m) = \sum_{i>m} P_\theta\left(i \in X^n\right)$, (b) is because for $i > m = n^\rho$, $\log i > \rho \log n$, and (c) follows from (61). Given $\rho > \varepsilon$ for an arbitrary *fixed* $\varepsilon > 0$, the resulting coefficient above is upper bounded by some constant κ.

Summing up the contributions of the terms of (58) from (28), (55), and (63), absorbing low order terms in a leading ε', we obtain that for $m = o(n^{1/3})$,

$$nR_n\left(L^*, \boldsymbol{\theta}\right) \leq \left(1+\varepsilon'\right)\frac{m-1}{2}\log\frac{n}{m^3} + \kappa n \sum_{i>m}\theta_i \log i. \tag{64}$$

For the second region, substituting $\alpha = 1/3$, and summing up the contributions of (60), (56), and (63) to (58), absorbing low order terms in ε', we obtain

$$nR_n\left(L^*, \boldsymbol{\theta}\right) \leq (1+\varepsilon')\frac{1}{2}\left(\rho+\frac{2}{3}\right)\left(\rho+\varepsilon'-\frac{1}{3}\right)(\log n)^2 n^{1/3} + \kappa n \sum_{i>m}\theta_i \log i. \tag{65}$$

Since (64)-(65) hold for every $m > n^\varepsilon$, there exists m^* for which the minimal bound is obtained. To bound the redundancy, we choose this m^*. Now, if the condition in (41) holds, then the second term in (64) and (65) is negligible w.r.t. the first term. Absorbing it in a leading ε, normalizing by n, yields the upper bound of (42), and concludes the proof of the Part I of Theorem 5.

For Part II of Theorem 5, we consider the bound of the second region in (65). If there exists $\rho^* = o\left(n^{1/3}/(\log n)\right)$ for which the condition in (43) holds, then both terms of (65) are of $o(n)$, yielding a total redundancy per symbol of $o(1)$. The proof of Theorem 5 is concluded. □

To prove Corollary 1, we use Wyner's inequality [32], which implies that for a finite entropy monotonic distribution,

$$\sum_{i\geq 1}\theta_i \log i = E_\theta\left[\log X\right] \leq H_\theta\left[X\right]. \tag{66}$$

Since the sum on the left hand side of (66) is finite if $H_\theta[X]$ is finite, there must exist some n_0 such that $\sum_{i>n_0}\theta_i \log i = o(1)$. Let $n > n_0$, then for $m^* = n$ and $\rho^* = 1$, condition (43) is satisfied. Therefore, (44) holds, and the proof of Corollary 1 is concluded. □

We now consider only the upper region in (58) with parameters α and ρ taking any valid value. (The code leading to the bound of the upper region can be applied even if the actual effective alphabet size is in the lower region.) We can sum up the contributions of (60), (56), and (63) to (58), absorbing low order terms in ε. Equation (56) is valid without the middle ε term as long as $\rho \geq \alpha + \varepsilon$. Since, in the upper region of m, $i \geq m$ is large enough, Elias' code for the integers can be used costing $(1+\varepsilon)\log i$ to code i, with $\varepsilon > 0$ which can be made arbitrarily small. Hence, the leading coefficient of the bound in (63) can be replaced by $(1+\varepsilon)(1+1/\rho)$. This yields the expression bounding the redundancy in (45). This expression applies to every valid choice of α and ρ, including the choice that minimizes the expression. Thus the proof of Theorem 6 is concluded. □

5.2 Examples

We demonstrate the use of the bounds of Theorems 5 and 6 with three typical distributions over the integers. We specifically show that the redundancy rate of $O\left(n^{1/3+\varepsilon}\right)$ bits overall is achievable when coding many of the typical monotonic distributions, and, in fact, for many distributions faster convergence rates are achievable with the codes provided in proving the theorems above. The assumption that very few unlikely symbols are likely to appear in a sequence generated by a monotonic distribution, which is reflected in the conditions in (41) and (43), is very realistic even in practical examples. Specifically, in the phone book example, there may be many rare names, but only very few of them may occur in a certain city, and the more common names constitute most of any possible phone book sequence.

5.2.1 Fast Decaying Distributions Over the Integers

Consider the monotonic distributions over the integers of the form,

$$\theta_i = \frac{a}{i^{1+\gamma}}, \; i = 1, 2, \ldots, \tag{67}$$

where $\gamma > 0$, and a is a normalization coefficient that guarantees that the probabilities over all integers sum to 1. It is easy to show by approximating summation by integration that for some $m \to \infty$,

$$\begin{aligned} S_m &\leq (1+\varepsilon)\frac{a}{\gamma m^{\gamma}} && (68) \\ \sum_{i>m} \theta_i \log i &\leq (1+\varepsilon)\frac{a \log m}{\gamma m^{\gamma}}. && (69) \end{aligned}$$

For $m = n^\rho$ and fixed ρ, the sum in (41) is thus $O\left(n^{1-\rho\gamma}\log n\right)$, which is $o\left(n^{1/3}(\log n)^2\right)$ for every $\rho \geq 2/(3\gamma)$. Specifically, as long as $\gamma \leq 2$ (slow decay), the minimal value of ρ required to guarantee negligibility of the sum in (41) is greater than 1/3. Using Theorem 5, this implies that for $\gamma \leq 2$, the second (upper) region of the upper bound in (42) holds with the minimal choice of $\rho^* = 2/(3\gamma)$. Plugging in this value in the second region of (40) (i.e., in (42)) yields the upper bound shown below for this region. For $\gamma > 2$, $2/(3\gamma) < 1/3$. Hence, (41) holds for $m^* = o\left(n^{1/3}\right)$. This means that for the distribution in (67) with $\gamma > 2$, the effective alphabet size is $o\left(n^{1/3}\right)$, and thus the achievable redundancy is in the first region of the bound of (42). Thus, even though the distribution is over an infinite alphabet, its compressibility behavior is similar to a distribution over a relatively small alphabet. To find the exact redundancy rate, we balance between the contributions of (55) and (63) in (58). As long as $1 - \rho\gamma < \rho$, condition (41) holds, and the contribution of small letters in (63) is negligible w.r.t. the other terms of the redundancy. Equality, implying $\rho^* = 1/(1+\gamma)$, achieves the minimal redundancy rate. Thus, for $\gamma > 2$,

$$\begin{aligned} nR_n\left(L^*, \boldsymbol{\theta}\right) &\overset{(a)}{\leq} (1+\varepsilon)\left[\frac{a(2\rho^*+1)}{\gamma} n^{1-\rho^*\gamma}\log n + \frac{n^{\rho^*}}{2}\left(1-3\rho^*\right)\log n\right] \\ &\overset{(b)}{=} (1+\varepsilon)\left(\frac{a^{\frac{3+\gamma}{1+\gamma}}}{\gamma} + \frac{1-\frac{3}{1+\gamma}}{2}\right) n^{\frac{1}{1+\gamma}}\log n \end{aligned} \tag{70}$$

where the first term in (a) follows from the bounds in (63) and (69), with $m = n^{\rho^*}$, and the second term from that in (55), and (b) follows from $\rho^* = 1/(1+\gamma)$. Note that for a fixed ρ^*, the factor 3 in the first term can be reduced to 2 with Elias' coding for the integers. The results described are summarized in the following corollary:

Corollary 2 *Let $\boldsymbol{\theta} \in \mathcal{M}$ be defined in (67). Then, there exists a universal code with length function $L^*(\cdot)$ that has only prior knowledge that $\boldsymbol{\theta} \in \mathcal{M}$, that can achieve universal coding redundancy*

$$R_n\left(L^*, \boldsymbol{\theta}\right) \leq \begin{cases} (1+\varepsilon)\frac{1}{9}\left(1+\frac{1}{\gamma}\right)\left(\frac{2}{\gamma}+\varepsilon-1\right)\frac{n^{1/3}(\log n)^2}{n}, & \textit{for } \gamma \leq 2, \\ (1+\varepsilon)\left(\frac{a^{\frac{3+\gamma}{1+\gamma}}}{\gamma} + \frac{1-\frac{3}{1+\gamma}}{2}\right)\frac{n^{\frac{1}{1+\gamma}}\log n}{n}, & \textit{for } \gamma > 2. \end{cases} \tag{71}$$

Corollary 2 gives the redundancy rates for all distributions defined in (67). For example, if $\gamma = 1$, the redundancy is $O\left(n^{1/3}(\log n)^2\right)$ bits overall with coefficient 2/9. For $\gamma = 3$, $O(n^{1/4}\log n)$ bits are required. For faster decays (greater γ) even smaller redundancy rates are achievable.

5.2.2 Geometric Distributions

Geometric distributions given by

$$\theta_i = p\left(1-p\right)^{i-1}; \quad i = 1, 2, \ldots, \tag{72}$$

where $0 < p < 1$, decay even faster than the distribution over the integers in (67). Thus their effective alphabet sizes are even smaller. This implies that a universal code can have even smaller redundancy than that presented in Corollary 2 when coding sequences generated by a geometric distribution (even if this is unknown in advance, and the only prior knowledge is that $\boldsymbol{\theta} \in \mathcal{M}$). Choosing $m = \ell \cdot \log n$, the contribution of low probability symbols in (63) to (58) can be upper bounded by

$$\begin{aligned} 2n \sum_{i>m} \theta_i \left(\log i + \log n\right) &\overset{(a)}{\leq} 2n(1-p)^m \log n + O\left(n(1-p)^m \log m\right) \\ &\overset{(b)}{=} 2n^{1+\ell \log(1-p)}(\log n) + O\left(n^{1+\ell \log(1-p)} \log \log n\right) \end{aligned} \tag{73}$$

where (a) follows from computing S_m using geometric series, and bounding the second term, and (b) follows from substituting $m = \ell \log n$ and representing $(1-p)^{\ell \log n}$ as $n^{\ell \log(1-p)}$. As long as $\ell \geq 1/(-\log(1-p))$, the expression in (73) is $O(\log n)$, thus negligible w.r.t. the redundancy upper bound of (42) with $m^* = \ell^* \log n = (\log n)/(-\log(1-p))$. Substituting this m^* in (42), we obtain the following corollary:

Corollary 3 *Let $\boldsymbol{\theta} \in \mathcal{M}$ be a geometric distribution defined in (72). Then, there exists a universal code with length function $L^*(\cdot)$ that has only prior knowledge that $\boldsymbol{\theta} \in \mathcal{M}$, that can achieve universal coding redundancy*

$$R_n\left(L^*, \boldsymbol{\theta}\right) \leq \frac{1+\varepsilon}{-2\log(1-p)} \cdot \frac{(\log n)^2}{n}. \tag{74}$$

Corollary 3 shows that if $\boldsymbol{\theta}$ parameterizes a geometric distribution, sequences governed by $\boldsymbol{\theta}$ can be coded with average universal coding redundancy of $O\left((\log n)^2\right)$ bits. Their effective alphabet size is $O(\log n)$, implying that larger symbols are very unlikely to occur. For example, for $p = 0.5$, the effective alphabet size is $\log n$, and $0.5(\log n)^2$ bits are required for a universal code. For $p = 0.75$, the effective alphabet size is $(\log n)/2$, and $(\log n)^2/4$ bits are required by a universal code.

5.2.3 Slow Decaying Distributions Over the Integers

Up to now, we considered fast decaying distributions, which all achieved the $O(n^{1/3+\varepsilon}/n)$ redundancy rate. We now consider a slowly decaying monotonic distribution over the integers, given by

$$\theta_i = \frac{a}{i\left(\log i\right)^{2+\gamma}},\ i = 2, 3, \ldots, \tag{75}$$

where $\gamma > 0$ and a is a normalizing factor (see, e.g., [12], [27]). This distribution has finite entropy only if $\gamma > 0$ (but is a valid infinite entropy distribution for $\gamma > -1$). Unlike the previous distributions, we need to use Theorem 6 to bound the redundancy for coding sequences generated by this distribution. Approximating the sum with an integral, the order of the third term of (45) is

$$n \sum_{i>m} \theta_i \log i = O\left(\frac{n}{(\log m)^\gamma}\right). \tag{76}$$

In order to minimize the redundancy bound of (45), we define $\rho = n^\ell$. For the minimum rate, all terms of (45) must be balanced. To achieve that, we must have

$$\alpha + 2\ell = 1 - 2\alpha = 1 - \gamma\ell. \tag{77}$$

The solution is $\alpha = \gamma/(4+3\gamma)$, and $\ell = 2/(4+3\gamma)$. Substituting these values in the expression of (45), with $\rho = n^\ell$, results in the first term in (45) dominating, and yields the following corollary:

Corollary 4 *Let $\boldsymbol{\theta} \in \mathcal{M}$ be defined in (75) with $\gamma > 0$. Then, there exists a universal code with length function $L^*(\cdot)$ that has only prior knowledge that $\boldsymbol{\theta} \in \mathcal{M}$, that can achieve universal coding redundancy*

$$R_n\left(L^*, \boldsymbol{\theta}\right) \leq (1+\varepsilon) \frac{n^{\frac{\gamma+4}{3\gamma+4}} (\log n)^2}{2n}. \tag{78}$$

Due to the slow decay rate of the distribution in (75), the effective alphabet size is much greater here. For $\gamma = 1$, for example, it is $n^{n^{2/7}}$. This implies that very large symbols are likely to appear in x^n. As γ increases though, the effective alphabet size decreases, and as $\gamma \to \infty$, $m \to n$. The redundancy rate increases due to the slow decay. For $\gamma \geq 1$, it is $O\left(n^{5/7}(\log n)^2/n\right)$. As $\gamma \to \infty$, since the distribution tends to decay faster, the redundancy rate tends to the finite alphabet rate of $O\left(n^{1/3}(\log n)^2/n\right)$. However, as the decay rate is slower $\gamma \to 0$, a non-diminishing redundancy rate is approached. Note that the proof of Theorem 6 does not limit the distribution to a finite entropy one. Therefore, the bound of (78) applies, in fact, also to $-1 < \gamma \leq 0$. However, for $\gamma \leq 0$, the per-symbol redundancy is no long diminishing.

6 Individual Sequences

In this section, we first show that individual sequences whose empirical distributions obey the monotonicity constraints can be universally compressed as well as the average case. We then study compression of sequences whose empirical distributions may diverge from monotonic. We demonstrate that under mild conditions, similar in nature to those of Theorems 5 and 6, redundancy that diminishes (slower than in the average case) w.r.t. the monotonic ML description length can be obtained. However, these results are only useful when the monotonic ML description length diverges only slightly from the (standard) ML description length of a sequence, i.e., the empirical distribution of a sequence only mildly violates monotonicity. Otherwise, the penalty of using an incorrect monotone model overwhelms the redundancy gain. We begin with sequences that obey the monotonicity constraints.

Theorem 7 *Fix an arbitrarily small $\varepsilon > 0$, and let $n \to \infty$. Let x^n be a sequence for which $\hat{\boldsymbol{\theta}} \in \mathcal{M}$, i.e., $\hat{\theta}_1 \geq \hat{\theta}_2 \geq \ldots$. Let $k = \hat{k}$ be the number of letters occurring in x^n. Then, there exists a code $L^*(\cdot)$ that achieves individual sequence redundancy w.r.t. $\hat{\boldsymbol{\theta}}_{\mathcal{M}} = \hat{\boldsymbol{\theta}}$ for x^n which is upper bounded by*

$$\hat{R}_n\left(L^*, x^n\right) \leq \begin{cases} (1+\varepsilon)\frac{k-1}{2n}\log\frac{n(\log n)^2}{k^3}, & \text{for } k \leq n^{1/3}, \\ (1+\varepsilon)(\log n)\left(\log\frac{k}{n^{1/3-\varepsilon}}\right)\frac{n^{1/3}}{n}, & \text{for } n^{1/3} < k = o(n), \\ (1+\varepsilon)\frac{1}{3}(\log n)^2\frac{n^{1/3}}{n}, & \text{for } n^{1/3} < k = O(n). \end{cases} \tag{79}$$

Note that by the monotonicity constraint, the number of symbols $\hat{k}$ occurring in x^n also equals to the maximal symbol in x^n. Since, in the individual sequence case, this maximal symbol defines the class considered and also to be consistent with Theorem 3, we use k to characterize the alphabet size of a given sequence. (The maximal symbol in the individual sequence case is equivalent to the alphabet size in the average case.) Finally, since $\hat{\boldsymbol{\theta}}$ is monotonic, $\hat{\boldsymbol{\theta}}_{\mathcal{M}} = \hat{\boldsymbol{\theta}}$.

Proof of Theorem 7: The result in Theorem 7 follows directly from the proof of Theorem 4. Both regions of the proof apply here, where instead of quantizing $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$, we quantize $\hat{\boldsymbol{\theta}}$ to $\hat{\boldsymbol{\theta}}'$ in a similar manner, and do not need to average over all sequences. In fact, instead of using any general $\hat{\boldsymbol{\varphi}}$ to code x^n, we can use $\hat{\boldsymbol{\theta}}'$ without any additional optimizations, where $\log n$ bits describe k. The description costs of $\hat{\boldsymbol{\theta}}'$ are almost the same as those of $\boldsymbol{\theta}'$. The factor 2 reduction in the last region is because it is sufficient here to replace n^2 by n in the denominators of (32). This is because for every occurring symbol $\hat{\theta}'_i \geq 1/n$ and $\delta_i \leq 1/n$, thus the first term of step (a) in (39) holds with the new grid, and B_2 in (36) reduces by a factor of 2. The quantization costs bounded in (28) and (39) are thus bounded similarly, where $\hat{\boldsymbol{\theta}}$ replaces $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}'$ replaces $\boldsymbol{\theta}'$. This results in the bounds in (79) and concludes the proof of Theorem 7. □

If one *a-priori* knows that x^n is likely to have been generated by a monotonic distribution, the case considered in Theorem 7 is with high probability the typical one. However, a typical sequence can also be one for which $\hat{\boldsymbol{\theta}} \notin \mathcal{M}$, where $\hat{\boldsymbol{\theta}}$ mildly violates the monotonicity. In the pure individual sequence setting (where no underlying distribution is assumed but some monotonicity assumption is reasonable for the empirical distribution of x^n), one can still observe sequences that have empirical distributions that are either monotonic or slightly diverge from monotonic. Coding for this more general case can apply the methods described in Section 5 to the individual sequence case. If the divergence from monotonicity is small, one may still achieve bounds of the same order of those presented in Theorem 7 with additional negligible cost of relaying which symbols are out of order. The next theorem, however, provides a general upper bound in the form of the bounds of Theorems 5 and 6 for the individual sequence redundancy w.r.t. the monotonic ML description length, as defined in (10). We begin, again, with some notation.

Recall the definition of an effective alphabet size $m \stackrel{\triangle}{=} m_\rho \stackrel{\triangle}{=} n^\rho$ (where $\rho = (\log m)/(\log n)$.) Now, use this definition for a specific individual sequence x^n. Let

$$\hat{\mathcal{R}}_n(m) \stackrel{\triangle}{=} \begin{cases} \frac{m-1}{2} \log \frac{n}{m}, & m \leq n^{1/3}, \\ m \log \frac{n}{m^2}, & n^{1/3} < m = o\left(\sqrt{n}\right), \\ \min_{\alpha<\rho} \left\{ \frac{\rho+1+\alpha}{2} \left(\rho - \alpha\right) \left(\log n\right)^2 n^\alpha + 3(\log e) n^{1-\alpha} \right\}, & \text{otherwise.} \end{cases} \quad (80)$$

Theorem 8 *Fix an arbitrarily small $\varepsilon > 0$, and let $n \to \infty$. Then, there exists a code with length function $L^*(\cdot)$, that achieves individual sequence redundancy w.r.t. the monotonic ML description length of x^n (as defined in (10)) bounded by*

$$\hat{R}_n\left(L^*, x^n\right) \leq \frac{1+\varepsilon}{n} \min_\rho \left\{ \hat{\mathcal{R}}_n\left(n^\rho\right) + \left(1 + \frac{1}{\rho}\right) \sum_{i>n^\rho, i \in x^n} \log i \right\} \quad (81)$$

for every x^n.

Theorem 8 shows that if one can find a relatively small effective alphabet of the symbols that occur in x^n, and the symbols outside this alphabet are small enough, x^n can be described with diminishing per-symbol redundancy w.r.t. its monotonic ML description length. This implies that as long as the occurring symbols are not too large, there exist a universal code w.r.t. a monotonic ML distribution for any such sequence x^n. This is unlike standard individual sequence compression w.r.t. the i.i.d. ML description length. Specifically, if the effective alphabet size is $O(n)$, and only a small number of symbols which are only polynomial in n occur, the universality cost is $O(\sqrt{n}(\log n)^2)$ bits overall, which gives diminishing per-symbol redundancy of $O((\log n)^2/\sqrt{n})$. This redundancy is much better than what can be achieved in standard compression. The penalty, of course, is when the empirical distribution of an individual sequence diverges significantly away from a monotonic one. While the monotonic redundancy can be made diminishing under mild conditions, there is a non-diminishing divergence cost by using the monotonic ML description length instead of the ML description length in that case. This implies that one should compress a sequence as generated by a monotonic distribution only if the total description length required to code x^n as such is shorter than the total description length required to code x^n with standard methods. As shown in the proof of Theorem 8, one prefix bit can inform the decoder which type of description is used.

Theorem 8 shows that as long as the effective alphabet size is polynomial in n, $\alpha = 0.5$ optimizes the third region of the upper bound, thus yielding the rate shown above, unless very large symbols occur in x^n. For small effective alphabets (the first region), there is no redundancy gain in using the monotonic ML description length over the ML description length. The reason, again, is that the bound is obtained for cases where the actual empirical distribution of a sequence may not be monotonic. One can still use an i.i.d. ML estimate w.r.t. only the effective alphabet, if the additional cost of symbols outside this alphabet is negligible, to better code such sequences. Theorem 8 also shows that if a very large symbol, such as $i = a^n$; $a > 1$, occurs in x^n, x^n cannot be universally compressed even w.r.t. its monotonic ML description length. This is because it is impossible to avoid the cost of $(1+\varepsilon)\log i = (1+\varepsilon)n \log a$ bits to describe this symbol to the decoder. The bound above and its proof below give a very powerful method to individually compress sequences that have an almost monotonic empirical distribution but may have some limited disorder, for which the monotonic ML description length diverges only negligibly from the ML description length.

Proof of Theorem 8: The proof follows the same steps as the proof of Theorems 5 and 6. Each value of m is tested and the best one is chosen, where the same coding costs described in the mentioned proof are computed for each m. In addition, one can test the cost of coding x^n using the description lengths for both $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_{\mathcal{M}}$. Then, one bit can be used to relay which ML estimator is used. If $\hat{\boldsymbol{\theta}}$ is used, the codes for coding individual sequences over large alphabets in either [21] or [25] can be used. In the first region in (81), the bound in [25] is obtained since $\log P_{\hat{\theta}}\left(x^n\right) \geq \log P_{\hat{\theta}_{\mathcal{M}}}\left(x^n\right)$ for every x^n. This bound yields smaller redundancy for this region than that obtained using $\hat{\boldsymbol{\theta}}_{\mathcal{M}}$ if $\hat{\boldsymbol{\theta}}_{\mathcal{M}} \neq \hat{\boldsymbol{\theta}}$. It implies that for small alphabets, if x^n does not have an empirical monotonic distribution, it is better coded, even in terms of universal coding redundancy, using standard universal compression methods without taking advantage of a monotonicity assumption.

For the other two regions, we start with a lemma.

Lemma 6.1 *Let* $\hat{\boldsymbol{\theta}}_{\mathcal{M}} = \left(\hat{\theta}_{1,\mathcal{M}}, \hat{\theta}_{2,\mathcal{M}}, \ldots, \hat{\theta}_{k,\mathcal{M}}\right)$ *be the monotonic ML estimator of* $\boldsymbol{\theta}$ *from* x^n, *i.e.,*

$\hat{\theta}_{1,\mathcal{M}} \geq \hat{\theta}_{2,\mathcal{M}} \geq \cdots \geq \hat{\theta}_{k,\mathcal{M}}$, *where* $k = \max\{x_1, x_2, \ldots, x_n\}$. *Then,*

$$\hat{\theta}_{k,\mathcal{M}} \geq \frac{1}{kn}. \tag{82}$$

Lemma 6.1 provides a lower bound on the minimal nonzero probability component of the monotonic ML estimator. This bound helps in designing the grid of points used to quantize the monotonic ML distribution of x^n, while maintaining bounded quantization costs. The proof of Lemma 6.1 is in Appendix C.

For m in the second region, we cannot use the grid in (18). The reason is that, here, the quantization cost is affected by both $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_{\mathcal{M}}$. This is unlike the average case, where the average respective vectors merge. To limit the quantization cost for very small probabilities, using Lemma 6.1, the minimal grid point must be $1/n^2$ or smaller. To make the quantization cost negligible w.r.t. the cost of describing the quantized ML, the ratio $\Delta_j/\varphi_{i,\mathcal{M}}$ between the spacing in interval j, and a quantized version $\varphi_{i,\mathcal{M}}$ of $\hat{\theta}_{i,\mathcal{M}}$ in the jth interval, must be $O(m/n)$. Hence, using the same methodology of the proof of Theorems 5 and 6, we define the jth interval for an effective alphabet $m = n^\rho = o\left(\sqrt{n}\right)$ as

$$\hat{I}_j = \left[\frac{n^{(j-1)\beta}}{n^2}, \frac{n^{j\beta}}{n^2}\right), \quad 1 \leq j \leq \hat{J}_\rho. \tag{83}$$

The spacing in the jth interval is

$$\hat{\Delta}_j^{(\rho)} = \frac{mn^{j\beta}}{n^3}. \tag{84}$$

This gives a total of

$$\hat{B}_\rho \leq \frac{n}{m}\log n \tag{85}$$

quantization points. Using the same methodology as in (21), this yields a representation cost of

$$L_R\left(\boldsymbol{\varphi}_m\right) \leq (1+\varepsilon)\, m \log \frac{n}{m^2} \tag{86}$$

where $\boldsymbol{\varphi}_m$ is the quantized version of $\hat{\boldsymbol{\theta}}_{\mathcal{M}}$ in which only the first m components of $\hat{\boldsymbol{\theta}}_{\mathcal{M}}$ are considered. Using the quantization with the grid defined in (83)-(86) in a code similar to the one used in the proof of Theorems 5 and 6, the *individual* quantization cost is given by

$$\begin{aligned}
\log \frac{P_{\hat{\theta}_{\mathcal{M}}}\left(x^n\right)}{P\left(x^n | m, S'_m, \boldsymbol{\varphi}_m\right)} &\overset{(a)}{\leq} n \sum_{i=1}^{m} \hat{\theta}_i \log \frac{\hat{\theta}_{i,\mathcal{M}}}{\varphi_{i,m}} + \log e \\
&\overset{(b)}{\leq} n (\log e) \sum_{i=1}^{m} \hat{\theta}_i \left| \frac{\delta_i}{\varphi_{i,m}} \right| + \log e \\
&\overset{(c)}{\leq} (\log e) \cdot \frac{n}{n^2} \cdot mn + (\log e) \cdot n \cdot \frac{mn^{j\beta}}{n^3} \cdot \frac{2n^2}{n^{j\beta}} + \log e \\
&= 3m(\log e) + \log e.
\end{aligned} \tag{87}$$

where (a) follows the same steps as in (60), (b) follows from $\ln(1+x) \leq x$, and then $x \leq |x|$, where $\delta_i \stackrel{\triangle}{=} \hat{\theta}_{i,\mathcal{M}} - \varphi_{i,m}$, and (c) follows from Lemma 6.1 and the definition of $\hat{I}_j$ in (83) (for the worst case first term, $|\delta_i| \leq 1/n^2$ and $\varphi_{i,m} \geq 1/(mn)$), from (84) and (83) (the second term), and since $\sum \hat{\theta}_i = 1$. The only additional non-negligible cost of coding sequences using a code as defined in the proof of Theorems 5 and 6 for a given m is the cost of coding all symbols $i > m$ that occur in x^n. Using a similar derivation to (54), with Elias' asymptotic code for the integers, this yields an additional cost of $(1+\varepsilon)(1+1/\rho)\sum_{i>n^\rho, i \in x^n} \log i$ code bits. Combining all costs, absorbing low order terms in ε, and normalizing by n, yields the second region of the bound in (81). Note that this bound also applies to the first region, but in that region, a tighter bound is obtained by using a code that uses the standard i.i.d. ML estimator $\hat{\boldsymbol{\theta}}$. This is because very fine quantization is needed to offset the cost of mismatch between $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_{\mathcal{M}}$. This quantization requires higher description costs than the description of a quantized *type* of a sequence when using standard compression. (This is not the case when $\hat{\boldsymbol{\theta}}$ obeys the monotonicity, as in Theorem 7. Even if $\hat{\boldsymbol{\theta}}$ does not obey monotonicity in the upper regions of the bound, this is not the case.)

For the last region of the bound, we follow the same steps above as was done for the upper region of the bound in Theorem 5 with a parameter α. The intervals are chosen, again, to guarantee bounded quantization costs. Hence,

$$\hat{I}_j = \left[\frac{n^{(j-1)\beta}}{n^{\rho+1+\alpha}}, \frac{n^{j\beta}}{n^{\rho+1+\alpha}}\right), \quad 1 \leq j \leq \hat{J}_\rho. \tag{88}$$

The spacing in the jth interval is

$$\hat{\Delta}_j^{(\rho)} = \frac{n^{j\beta}}{n^{\rho+1+2\alpha}}. \tag{89}$$

This gives a total of

$$\hat{B}_\rho \leq 0.5 n^\alpha \lceil (\rho+1+\alpha)\log n \rceil \tag{90}$$

quantization points. Using the same methodology as in (56), this yields a representation cost of

$$L_R(\boldsymbol{\varphi}_m) \leq (1+\varepsilon)\frac{\rho+1+\alpha}{2}(\rho+\varepsilon-\alpha)(\log n)^2 n^\alpha. \tag{91}$$

Similarly to (87),

$$\log\frac{P_{\hat{\theta}_{\mathcal{M}}}(x^n)}{P(x^n | m, S'_m, \boldsymbol{\varphi}_m)} \stackrel{(a)}{\leq} (\log e)\frac{n^{\rho+2}}{n^{\rho+1+\alpha}} + (\log e) 2n^{1-\alpha} + \log e = 3(\log e)n^{1-\alpha} + \log e \tag{92}$$

where (a) follows from similar steps to (a)-(c) of (87). Using Lemma 6.1, $\varphi_{i,m} \geq 1/n^{\rho+1}$ and $|\delta_i| \leq 1/n^{\rho+1+\alpha}$, leading to the first term. Bounding $|\delta_i| \leq n^{j\beta}/n^{\rho+1+2\alpha}$ and $\varphi_{i,m} \geq n^{(j-1)\beta}/n^{\rho+1+\alpha}$ leads to the second term. Note that as before, m is used here in place of k, because using an effective alphabet m, all greater symbols are packed together as one symbol, and the additional cost to describe them is reflected in an additional term. Adding this additional term with an identical expression to that in the lower regions, absorbing low order terms in ε, and normalizing by n, yields the third region of the bound in (81). Since the bound holds for every α and every $\rho > \alpha$, it can be optimized to give the values that attain the minimum, concluding the proof of Theorem 8. □

7 Summary and Conclusions

Universal compression of sequences generated by monotonic distributions was studied. We showed that for finite alphabets, if one has the prior knowledge of the monotonicity of a distribution, one can reduce the cost of universality. For alphabets of $o(n^{1/3})$ letters, this cost reduces from $0.5\log(n/k)$ bits per each unknown probability parameter to $0.5\log(n/k^3)$ bits per each unknown probability parameter. Otherwise, for alphabets of $O(n)$ letters, one can compress such sources with overall redundancy of $O(n^{1/3+\varepsilon})$ bits. This is a significant decrease in redundancy from $O(k\log n)$ or $O(n)$ bits overall that can be achieved if no side information is available about the source distribution. Redundancy of $O(n^{1/3+\varepsilon})$ bits overall can also be achieved for much larger alphabets including infinite alphabets for fast decaying monotonic distributions. Sequences generated by slower decaying distributions can also be compressed with diminishing per-symbol redundancy costs under some mild conditions and specifically if they have finite entropy rates. Examples for well-known monotonic distributions demonstrated how the diminishing redundancy decay rates can be computed by applying the bounds that were derived. Finally, the average case results were extended to individual sequences. Similar convergence rates were shown for sequences that have empirical monotonic distributions. Furthermore, universal redundancy bounds w.r.t. the monotonic ML description length of a sequence were also derived for the more general case. Under some mild conditions, these bounds still exhibit diminishing per-symbol redundancies.

Appendix A – Proof of Theorem 1

The proof follows the same steps used in [25] and [26] to lower bound the maximin redundancies for large alphabets and patterns, respectively, using the weak version of the *redundancy-capacity theorem* [5]. This version ties between the maximin universal coding redundancy and the capacity of a channel defined by the conditional probability $P_\theta(x^n)$. We define a set $\boldsymbol{\Omega}_{\mathcal{M}_k}$ of points $\boldsymbol{\theta} \in \mathcal{M}_k$. Then, show that these points are *distinguishable* by observing X^n, i.e., the probability that X^n

generated by $\boldsymbol{\theta} \in \boldsymbol{\Omega}_{\mathcal{M}_k}$ appears to have been generated by another point $\boldsymbol{\theta}' \in \boldsymbol{\Omega}_{\mathcal{M}_k}$ diminishes with n. Then, using Fano's inequality [3], the number of such distinguishable points is a lower bound on $R_n^-(\mathcal{M}_k)$. Since $R_n^+(\mathcal{M}_k) \geq R_n^-(\mathcal{M}_k)$, it is also a lower bound on the average minimax redundancy. The two regions in (6) result from a threshold phenomenon, where there exists a value k_m of k that maximizes the lower bound, and can be applied to all $\mathcal{M}_k$ for $k \geq k_m$.

We begin with defining $\boldsymbol{\Omega}_{\mathcal{M}_k}$. Let $\boldsymbol{\omega}$ be a vector of grid components, such that the last $k-1$ components $\theta_i,\ i = 2, \ldots, k$, of $\boldsymbol{\theta} \in \boldsymbol{\Omega}_{\mathcal{M}_k}$ must satisfy $\theta_i \in \boldsymbol{\omega}$. Let ω_b be the bth point in $\boldsymbol{\omega}$, and define $\omega_0 = 0$ and

$$\omega_b \overset{\triangle}{=} \sum_{j=1}^{b} \frac{2(j-\frac{1}{2})}{n^{1-\varepsilon}} = \frac{b^2}{n^{1-\varepsilon}}, \quad b = 1, 2, \ldots. \tag{A.1}$$

Then, for the bth point in $\boldsymbol{\omega}$,

$$b = \sqrt{\omega_b} \cdot \sqrt{n}^{1-\varepsilon}. \tag{A.2}$$

To count the number of points in $\boldsymbol{\Omega}_{\mathcal{M}_k}$, let us first consider the standard i.i.d. case, where there is no monotonicity requirement, and count the number of points in $\boldsymbol{\Omega}$, which is defined similarly, but without the monotonicity requirement (i.e., $\boldsymbol{\Omega}_{\mathcal{M}_k} \subseteq \boldsymbol{\Omega}$). Let b_i be the index of θ_i in $\boldsymbol{\omega}$, i.e., $\theta_i = \omega_{b_i}$. Then, from (A.1)-(A.2) and since the components of $\boldsymbol{\theta}$ are probabilities,

$$\sum_{i=2}^{k} \frac{b_i^2}{n^{1-\varepsilon}} = \sum_{i=2}^{k} \omega_{b_i} = \sum_{i=2}^{k} \theta_i \leq 1. \tag{A.3}$$

It follows that for $\boldsymbol{\theta} \in \boldsymbol{\Omega}$,

$$\sum_{i=2}^{k} b_i^2 \leq n^{1-\varepsilon}. \tag{A.4}$$

Hence, since the components b_i are nonnegative integers,

$$\begin{aligned} M \overset{\triangle}{=} |\boldsymbol{\Omega}| \quad &\geq \quad \sum_{b_2=0}^{\left\lfloor \sqrt{n^{1-\varepsilon}} \right\rfloor} \ \sum_{b_3=0}^{\left\lfloor \sqrt{n^{1-\varepsilon}-b_2^2} \right\rfloor} \cdots \sum_{b_k=0}^{\left\lfloor \sqrt{n^{1-\varepsilon}-\sum_{i=2}^{k-1} b_i^2} \right\rfloor} 1 \\ &\overset{(a)}{\geq} \int_0^{\sqrt{n^{1-\varepsilon}}} \int_0^{\sqrt{n^{1-\varepsilon}-x_2^2}} \cdots \int_0^{\sqrt{n^{1-\varepsilon}-\sum_{i=2}^{k-1} x_i^2}} dx_k \cdots dx_3 dx_2 \ \overset{(b)}{\overset{\triangle}{=}} \ \frac{V_{k-1}\left(\sqrt{n}^{1-\varepsilon}\right)}{2^{k-1}} \end{aligned} \tag{A.5}$$

where $V_{k-1}\left(\sqrt{n}^{1-\varepsilon}\right)$ is the volume of a $k-1$ dimensional sphere with radius $\sqrt{n}^{1-\varepsilon}$, (a) follows from monotonic decrease of the function in the integrand for all integration arguments, and (b) follows since its left hand side computes the volume of the positive quadrant of this sphere. Note that this is a different proof from that used in [25]-[26] for this step. Applying the monotonicity constraint, all permutations of $\boldsymbol{\theta}$ that are not monotonic must be taken out of the grid. Hence,

$$M_{\mathcal{M}_k} \overset{\triangle}{=} |\boldsymbol{\Omega}_{\mathcal{M}_k}| \geq \frac{V_{k-1}\left(\sqrt{n}^{1-\varepsilon}\right)}{k! \cdot 2^{k-1}}, \tag{A.6}$$

where dividing by $k!$ is a worst case assumption, yielding a lower bound and not an equality. This leads to a lower bound equal to that obtained for patterns in [26] on the number of points in $\boldsymbol{\Omega}_{\mathcal{M}_k}$. Specifically, the bound achieves a maximal value for $k_m = \left(\pi n^{1-\varepsilon}/2\right)^{1/3}$ and then decreases to eventually become smaller than 1. However, for $k > k_m$, one can consider a monotonic distribution for which all components $\theta_i; i > k_m$, of $\boldsymbol{\theta}$ are zero, and use the bound for k_m.

Distinguishability of $\boldsymbol{\theta} \in \boldsymbol{\Omega}_{\mathcal{M}_k}$ is a direct result of distinguishability of $\boldsymbol{\theta} \in \boldsymbol{\Omega}$, which is shown in Lemma 3.1 in [25], i.e., there exits an estimator $\hat{\boldsymbol{\Theta}}_g(X^n) \in \boldsymbol{\Omega}$ for which the estimate $\hat{\boldsymbol{\theta}}_g$ satisfies $\lim_{n\to\infty} P_\theta\left(\hat{\boldsymbol{\theta}}_g \neq \boldsymbol{\theta}\right) = 0$ for all $\boldsymbol{\theta} \in \boldsymbol{\Omega}$. Since this is true for all points in $\boldsymbol{\Omega}$, it is also true for all points in $\boldsymbol{\Omega}_{\mathcal{M}_k} \subseteq \boldsymbol{\Omega}$, where now, $\hat{\boldsymbol{\theta}}_g \in \boldsymbol{\Omega}_{\mathcal{M}_k}$. Assuming all points in $\boldsymbol{\Omega}_{\mathcal{M}_k}$ are equally probable to generate X^n, we can define an average error probability $P_e \stackrel{\triangle}{=} \Pr\left[\hat{\boldsymbol{\Theta}}_g(X^n) \neq \boldsymbol{\Theta}\right] = \sum_{\boldsymbol{\theta}\in\boldsymbol{\Omega}_{\mathcal{M}_k}} P_\theta\left(\hat{\boldsymbol{\theta}}_g \neq \boldsymbol{\theta}\right)/M_{\mathcal{M}_k}$. Using the redundancy-capacity theorem,

$$\begin{aligned} nR_n^-\left[\mathcal{M}_k\right] &\geq C\left[\mathcal{M}_k \to X^n\right] \overset{(a)}{\geq} I[\boldsymbol{\Theta}; X^n] = H\left[\boldsymbol{\Theta}\right] - H\left[\boldsymbol{\Theta}|X^n\right] \\ &\overset{(b)}{=} \log M_{\mathcal{M}_k} - H\left[\boldsymbol{\Theta}|X^n\right] \overset{(c)}{\geq} \left(1-P_e\right)\left(\log M_{\mathcal{M}_k}\right) - 1 \\ &\overset{(d)}{\geq} \left(1-o(1)\right)\log M_{\mathcal{M}_k}, \end{aligned} \tag{A.7}$$

where $C\left[\mathcal{M}_k \to X^n\right]$ denotes the capacity of the respective channel and $I[\boldsymbol{\Theta}; X^n]$ is the mutual information induced by the joint distribution $\Pr\left(\Theta = \theta\right) \cdot P_\theta\left(X^n\right)$. Inequality (a) follows from the definition of capacity, equality (b) from the uniform distribution of $\boldsymbol{\Theta}$ in $\boldsymbol{\Omega}_{\mathcal{M}_k}$, inequality (c) from Fano's inequality, and (d) follows since $P_e \to 0$. Lower bounding the expression in (A.6) for the two regions (obtaining the same bounds as in [26]), then using (A.7), normalizing by n, and absorbing low order terms in ε, yields the two regions of the bound in (6). The proof of Theorem 1 is concluded. □

Appendix B – Proof of Theorem 2

To prove Theorem 2, we use the *random-coding* strong version of the redundancy-capacity theorem [17]. The idea is similar to the weak version used in Appendix A. We assume that grids $\boldsymbol{\Omega}_{\mathcal{M}_k}$ of points are uniformly distributed over $\mathcal{M}_k$, and one grid is selected randomly. Then, a point in the selected grid is randomly selected under a uniform prior to generate X^n. Showing distinguishability within a selected grid, for every possible random choice of $\boldsymbol{\Omega}_{\mathcal{M}_k}$, implies that a lower bound on the cardinality of $\boldsymbol{\Omega}_{\mathcal{M}_k}$ for every possible choice is essentially a lower bound on the overall sequence redundancy for most sources in $\mathcal{M}_k$.

The construction of $\mathbf{\Omega}_{\mathcal{M}_k}$ is identical to that used in [26] to construct a grid of sources that generate patterns. We pack spheres of radius $n^{-0.5(1-\varepsilon)}$ in the parameter space defining $\mathcal{M}_k$. The set $\mathbf{\Omega}_{\mathcal{M}_k}$ consists of the center points of the spheres. To cover the space $\mathcal{M}_k$, we randomly select a random shift of the whole lattice under a uniform distribution. The cardinality of $\mathbf{\Omega}_{\mathcal{M}_k}$ is lower bounded by the relation between the volume of $\mathcal{M}_k$, which equals (as shown in [26]) $1/[(k-1)!k!]$, and the volume of a single sphere, with factoring also of a packing density (see, e.g., [2]). This yields eq. (55) in [26],

$$M_{\mathcal{M}_k} \geq \frac{1}{(k-1)! \cdot k! \cdot V_{k-1}\left(n^{-0.5(1-\varepsilon)}\right) \cdot 2^{k-1}}, \tag{B.1}$$

where $V_{k-1}\left(n^{-0.5(1-\varepsilon)}\right)$ is the volume of a $k-1$ dimensional sphere with radius $n^{-0.5(1-\varepsilon)}$ (see, e.g., [2] for computation of this volume).

For distinguishability, it is sufficient to show that there exists an estimator $\hat{\mathbf{\Theta}}_g(X^n) \in \mathbf{\Omega}_{\mathcal{M}_k}$ such that $\lim_{n\to\infty} P_\Theta\left[\hat{\mathbf{\Theta}}_g(X^n) \neq \mathbf{\Theta}\right] = 0$ for every choice of $\mathbf{\Omega}_{\mathcal{M}_k}$ and for every choice of $\mathbf{\Theta} \in \mathbf{\Omega}_{\mathcal{M}_k}$. This is already shown in Lemma 4.1 in [25] for a larger grid $\mathbf{\Omega}$ of i.i.d. sources, which is constructed identically to $\mathbf{\Omega}_{\mathcal{M}_k}$ over the complete $k-1$ dimensional probability simplex. Therefore, by the monotonicity requirement, for every $\mathbf{\Omega}_{\mathcal{M}_k}$, there exists such $\mathbf{\Omega}$, such that $\mathbf{\Omega}_{\mathcal{M}_k} \subseteq \mathbf{\Omega}$. Since Lemma 4.1 in [25] holds for $\mathbf{\Omega}$, it then must also hold for the smaller grid $\mathbf{\Omega}_{\mathcal{M}_k}$. Note that distinguishability is easier to prove here than for patterns because $\hat{\mathbf{\Theta}}_g(X^n)$ is obtained directly form X^n and not from its pattern as in [26]. Now, since all the conditions of the strong random-coding version of the redundancy-capacity theorem hold, taking the logarithm of bound in (B.1), absorbing low order terms in ε, and normalizing by n, leads to the first region of the bound in (7). More detailed steps follow those found in [26].

The second region of the bound is handled in a manner related to the second region of the bound of Theorem 1. However, here, we cannot simply set the probability of all symbols $i > k_m$ to zero, because all possible valid sources must be included in one of the grids $\mathbf{\Omega}_{\mathcal{M}_k}$ to generate a complete covering of $\mathcal{M}_k$. As was done in [26], we include sources with $\theta_i > 0$ for $i > k_m$ in the grids $\mathbf{\Omega}_{\mathcal{M}_k}$, but do not include them in the lower bound on the number of grid points. Instead, for $k > k_m$, we bound the number of points in a k_m-dimensional cut of $\mathcal{M}_k$ for which the remaining $k - k_m$ components of $\boldsymbol{\theta}$ are very small (and insignificant). This analysis is valid also for $k > n$. Distinguishability for $k > k_m$ is shown for i.i.d. non-monotonically restricted distributions in the proof of Lemma 6.1 in [26]. As before, it carries over to monotonic distributions, since as before, for each $\mathbf{\Omega}_{\mathcal{M}_k}$, there exists an unrestricted corresponding $\mathbf{\Omega}$, such that $\mathbf{\Omega}_{\mathcal{M}_k} \subseteq \mathbf{\Omega}$. The choice of $k_m = 0.5(n^{1-\varepsilon}/\pi)^{1/3}$ gives the maximal bound w.r.t. k. Since, again, all conditions of the strong version of the redundancy-capacity theorem are satisfied, the second region of the bound is obtained. Again, more detailed steps can be found in [26]. This concludes the proof of Theorem 2. □

Appendix C – Proof of Lemma 6.1

For cardinality k, we consider the largest component of $\hat{\boldsymbol{\theta}}_{\mathcal{M}}$; $\hat{\theta}_{1,\mathcal{M}}$, as the constraint component, i.e., $\hat{\theta}_{1,\mathcal{M}} = 1 - \sum_{i=2}^{k} \hat{\theta}_{i,\mathcal{M}}$. For any given probability parameter $\boldsymbol{\varphi}$ of cardinality k with $\varphi_1 > 0$, we have

$$P_{\varphi}\left(x^n\right) = \varphi_1^{n_x(1)}\left(1-\varphi_1\right)^{n-n_x(1)} \cdot \prod_{i=2}^{k}\left(\frac{\varphi_i}{1-\varphi_1}\right)^{n_x(i)} \triangleq \varphi_1^{n_x(1)}\left(1-\varphi_1\right)^{n-n_x(1)} \prod_{i=2}^{k} \vartheta_i^{n_x(i)} \qquad \text{(C.1)}$$

where we recall that $n_x(i)$ is the occurrence count of i in x^n. Therefore, maximization of $P_{\varphi}\left(x^n\right)$ w.r.t. φ_1 is independent of the maximization over ϑ_i; $i > 1$, and is obtained for $\varphi_1 = \hat{\theta}_1 = n_x(1)/n$. Since for all $i > 1$, $\hat{\theta}_{1,\mathcal{M}} \geq \hat{\theta}_{i,\mathcal{M}}$, $\hat{\theta}_{1,\mathcal{M}}$ can thus only increase from $\hat{\theta}_1$ by the monotonicity constraint. (Note that the monotonicity constraint implies a water filling [3] optimization to achieve $\hat{\boldsymbol{\theta}}_{\mathcal{M}}$.) Hence, $\hat{\theta}_{1,\mathcal{M}} \geq n_x(1)/n$.

Now, using the result above, we show that the derivative of $\ln P_{\varphi_{\mathcal{M}}}\left(x^n\right)$ w.r.t. $\varphi_{k,\mathcal{M}}$ is positive for $\varphi_{k,\mathcal{M}} < 1/(kn)$ and a monotonic $\boldsymbol{\varphi}_{\mathcal{M}}$. A component of a parameter vector $\boldsymbol{\varphi}_{\mathcal{M}}$, which is monotonic, can be expressed as

$$\varphi_{i,\mathcal{M}} = \sum_{\ell=i}^{k} \varphi'_{\ell}, \quad \varphi'_{\ell} \geq 0. \qquad \text{(C.2)}$$

Hence,

$$\begin{aligned} \left.\frac{\partial \ln P_{\varphi_{\mathcal{M}}}\left(x^n\right)}{\partial \varphi_{k,\mathcal{M}}}\right|_{\varphi_{1,\mathcal{M}}=\hat{\theta}_{1,\mathcal{M}}} &\overset{(a)}{=} \left.\frac{\partial \ln P_{\varphi_{\mathcal{M}}}\left(x^n\right)}{\partial \varphi'_{k}}\right|_{\varphi_{1,\mathcal{M}}=\hat{\theta}_{1,\mathcal{M}}} \\ &\overset{(b)}{=} \sum_{i=2}^{k} \frac{n_x(i)}{\varphi_{i,\mathcal{M}}} - \frac{(k-1)n_x(1)}{\hat{\theta}_{1,\mathcal{M}}} \\ &\overset{(c)}{>} \frac{kn_x(k)}{\hat{\theta}_k} - \frac{kn_x(1)}{\hat{\theta}_1} \overset{(d)}{=} 0 \end{aligned} \qquad \text{(C.3)}$$

where (a) follows from $\varphi_{k,\mathcal{M}}$ being the smallest nonzero component of $\boldsymbol{\varphi}_{\mathcal{M}}$, (b) is since by (C.2), φ'_k is included in all terms, and

$$\varphi_{1,\mathcal{M}} = 1 - \sum_{i=2}^{k} \varphi_{i,\mathcal{M}} = 1 - \sum_{i=2}^{k-1}(i-1)\varphi'_i - (k-1)\varphi_{k,\mathcal{M}}, \qquad \text{(C.4)}$$

where the last equality follows from (C.2), (c) follows by omitting all terms of the sum except $i = k$, from the assumption that $\varphi_{k,\mathcal{M}} < 1/(nk) \leq \hat{\theta}_k/k$, and since $\hat{\theta}_{1,\mathcal{M}} \geq n_x(1)/n = \hat{\theta}_1$, and (d) follows since its left hand side is 0 for the (i.i.d.) ML parameter values. Hence, $P_{\varphi_{\mathcal{M}}}(x^n)$ must increase, with $\varphi_{1,\mathcal{M}}$ taking its optimal value, for all $\boldsymbol{\varphi}_{\mathcal{M}}$ for which $\varphi_{k,\mathcal{M}} < 1/(nk)$, and the maximum is thus achieved for $\hat{\theta}_{k,\mathcal{M}} \geq 1/(nk)$. □

References

[1] J. Åberg, Y. M. Shtarkov, and B. J. M. Smeets, "Multialphabet coding with separate alphabet description," in *Proceedings of Compression and Complexity of Sequences*, pp. 56-65, Jun. 1997.

[2] J. H. Conway, N. J. A. Sloane, *Sphere Packings, Lattices and Groups*, Springer-Verlag, Third Edition, 1998.

[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, second edition, John Wiley & Sons, 2006.

[4] I. Csiszar and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems.*, Academic Press, New York, 1981.

[5] L. D. Davisson, "Universal noiseless coding," *IEEE Trans. Inform. Theory*, vol. IT-19, no. 6, pp. 783-795, Nov. 1973.

[6] L. D. Davisson, and A. Leon-Garcia, "A source matching approach to finding minimax codes," *IEEE Trans. Inform. Theory*, vol. IT-26, no. 2, pp. 166-174, Mar. 1980.

[7] P. Elias, "Universal codeword sets and representation of the integers," *IEEE Trans. Inform. Theory*, vol. IT-21, no. 2, pp. 194-203, March 1975.

[8] B. M. Fitingof, "Optimal coding in the case of unknown and changing message statistics," *Probl. Inform. Transm.*, vol. 2, no. 2, pp. 1-7, 1966.

[9] B. M. Fitingof, "The compression of discrete information," *Probl. Inform. Transm.*, vol. 3, no. 3, pp. 22-29, 1967.

[10] D. P. Foster, R. A. Stine, and A. J. Wyner, "Universal codes for finite sequences of integers drawn from a monotone distribution," *IEEE Trans. Inform. Theory*, vol. 48, no. 6, pp. 1713-1720, June 2002.

[11] R. G. Gallager, "Source coding with side information and universal coding," unpublished manuscript, September 1976.

[12] G. M. Gemelos and T. Weissman, "On the entropy rate of pattern processes," *IEEE Trans. Inform. Theory*, vol. 52, no. 9, pp. 3994-4007, Sept. 2006.

[13] L. Györfi, I. Páli, and E. C. van der Meulen, "There is no universal code for an infinite source alphabet," *IEEE Trans. Inform. Theory*, vol. 40, no. 1, pp. 267-271, Jan. 1994.

[14] N. Jevtić, A. Orlitsky, and N. P. Santhanam, "A lower bound on compression of unknown alphabets," *Theoret. Comput. Sci.*, vol. 332, no. 1-3, pp. 293-311, 2005.

[15] J. C. Kieffer, "A unified approach to weak universal source coding," *IEEE Trans. Inform. Theory*, vol. IT-24, no. 6, pp. 674-682, Nov. 1978.

[16] M. Khosravifard, H. Saidi, M. Esmaeili, and T. A. Gulliver, "The minimum average code for finite memoryless monotone sources," *IEEE Trans. Inform. Theory*, vol. 52, no. 3, pp. 955-975, Mar. 2007.

[17] N. Merhav and M. Feder, "A strong version of the redundancy-capacity theorem of universal coding," *IEEE Trans. Inform. Theory*, vol. no. 3, 41, pp. 714-722, May 1995.

[18] N. Merhav, G. Seroussi, and M. J. Weinberger, "Optimal prefix codes for sources with two-sided geometric distributions," *IEEE Trans. Inform. Theory*, vol. 46, no. 1, pp. 121-135, Jan. 2000.

[19] N. Merhav, G. Seroussi, and M. J. Weinberger, "Coding of sources with two-sided geometric distributions and unknown parameters," *IEEE Trans. Inform. Theory*, vol. 46, no. 1, pp. 229-236, Jan. 2000.

[20] A. Orlitsky, N. P. Santhanam, and J. Zhang, "Universal compression of memoryless sources over unknown alphabets," *IEEE Trans. Inform. Theory*, vol. 50, no. 7, pp. 1469-1481, July 2004.

[21] A. Orlitsky, and N. P. Santhanam, "Speaking of infinity," *IEEE Trans. Inform. Theory*, vol. 50, no. 10, pp. 2215-2230, Oct. 2004.

[22] J. Rissanen, “Minimax codes for finite alphabets,” *IEEE Trans. Inform. Theory*, vol. IT-24, no. 3, pp. 389-392, May 1978.

[23] J. Rissanen, “Universal coding, information, prediction, and estimation,” *IEEE Trans. Inform. Theory*, vol. IT-30, no. 4, pp. 629-636, Jul. 1984.

[24] B. Ya. Ryabko, “Coding of a source with unknown but ordered probabilities,” *Problems of Information Transmission*, vol. 15, no. 2, pp. 134-138, Oct. 1979.

[25] G. I. Shamir, “On the MDL principle for i.i.d. sources with large alphabets,” *IEEE Trans. Inform. Theory*, vol. 52, no. 5, pp. 1939-1955, May 2006.

[26] G. I. Shamir, “Universal lossless compression with unknown alphabets - the average case”, *IEEE Trans. Inform. Theory*, vol. 52, no. 11, pp. 4915-4944, Nov. 2006.

[27] G. I. Shamir, “Patterns of sequences and their entropy,” submitted to *IEEE Trans. Inform. Theory*. Also in *Arxiv:cs.IT/0605046*.

[28] G. I. Shamir, “A new redundancy bound for universal lossless compression of unknown alphabets,” in *Proceedings of The 38th Annual Conference on Information Sciences and Systems*, Princeton, New-Jersey, U.S.A., pp. 1175-1179, Mar. 17-19, 2004.

[29] Y. M. Shtarkov, “Universal sequential coding of single messages,” *Problems of Information Transmission*, 23(3):3-17, Jul.-Sep. 1987.

[30] L. R. Varshney and V. K. Goyal, “Ordered and disordered source coding,” in *Information Theory & Applications Workshop (ITA)*, San Diego, California, Feb. 6-10, 2006.

[31] L. R. Varshney and V. K. Goyal, “On universal coding of unordered data,” in *Information Theory & Applications Workshop (ITA)*, San Diego, California, Jan. 29-Feb. 2, 2007.

[32] A. D. Wyner, “An upper bound on the entropy series,” *Inform. Contr.*, vol. 20, pp. 176-181, 1972.